The Dissertation Committee for Sandra Elaine Pelc

certifies that this is the approved version of the following dissertation:

Determination of the genetic basis of seed oil composition and melting point—adaptive quantitative traits—and their fitness effects in *Arabidopsis thaliana*

Committee:

_____

Craig Randal Linder, Supervisor

_____

Thomas E. Juenger

_____

Robert K. Jansen

_____

Enamul Huq

_____

Mikhail V. Matz

_____

Claus O. Wilke

Determination of the genetic basis of seed oil composition and melting point—adaptive quantitative traits—and their fitness effects in *Arabidopsis thaliana*

by

Sandra Elaine Pelc, B.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2013

Dedication

This dissertation is dedicated to my grandmother Margaret (Peggy) Kane for making me the person that I am today and to my advisor Dr. Randy Linder for making me the scientist I am today.

Acknowledgments

I would like to express my most heartfelt gratitude to my advisor, Dr. Randy Linder, for encouraging me to further my education. His guidance and support has changed the course of my journey through life. I would also like to thank Dr. Beryl Simpson for taking a chance on hiring an inexperienced freshman as a lab assistant and thereby fostering my love of research. I would like to express my appreciation of my entire committee (Dr. Tom Juenger, Dr. Bob Jansen, Dr. Enamel Huq, Dr. Claus Wilke and Dr. Misha Matz) for their time, advice, and support throughout my graduate career.

The love and support of my wonderful family and friends has meant the world to me through the roller coaster of the last several years. I would especially like to thank my brother Aaron Reba and my friends Teofil Nakov, Ginnie Morrison, and Anushree Sanyal for the endless discussions of programming, statistics, biology, life, the universe and everything.

Determination of the genetic basis of seed oil composition and melting point—adaptive quantitative traits—and their fitness effects in *Arabidopsis thaliana*

Sandra Elaine Pelc, Ph.D.

The University of Texas at Austin, 2013

Supervisor: Craig Randal Linder

Evidence indicates seed oil melting point is likely an adaptive quantitative trait in many flowering plant species. An adaptive hypothesis suggests selection has changed the melting point of seed oils to covary with germination temperatures because of a trade-off between total energy stores and the rate of energy acquisition during germination under competition. The predicted differences in relative fitness under different temperatures have not yet been tested and little is known about the genetic basis of differences in oil composition. I used *Arabidopsis thaliana* to: (1) assess the fitness consequences of high and low melting point seeds germinating at different temperatures, (2) assess what genes underlie natural variation in seed oil composition, and (3) consider how these genes may be used to create oils with particular characteristics. To assess the effects of seed oil

melting point on timing of seedling emergence and fitness, I competed high and low melting point lines of *A. thaliana* under cold and warm temperatures. Emergence timing between these lines was not significantly different at either temperature, which comported with warm temperature predictions but not cold temperature predictions. Under all conditions, plants competing against high melting point lines had lower fitness relative to those against low melting point lines, which matched expectations for undifferentiated emergence times. To assess the genetic basis of naturally occurring variation in seed oil melting point, the seed oil compositions of 391 accessions of *A. thaliana* were used in a genome-wide association study. Twelve genes were tightly linked with SNPs significantly associated with seed oil melting point variation. Seven encoded lipid synthesis enzymes or regulatory products. The remaining 5 encoded products with no clear relation to seed oil melting point. Results suggest selection can alter quantitative trait variation in response to local conditions through a small set of genes. 268 seed-expressed, candidate genes were linked to 103 SNPs associated with *A. thaliana* seed oil fatty acids. Eight genes were involved in lipid metabolism, and thirty-four encoded regulatory products. I discuss how knowledge of these genes can be used to breed and engineer desirable oil compositions for industry and nutrition.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Quantitative traits result from the contribution of many genes. Variation in these traits results from variation in some subset of these genes and the influence of environmental variation. Understanding the genetic basis of adaptation of quantitative traits is an enduring goal of evolutionary biology with multiple aspects as of yet unresolved (reviewed in Orr, 2005).

Historically, studies attempting to identify the genes underlying quantitative genetic variation were limited by marker availability, leaving the genetics of adaptation of quantitative traits largely a black box. Researchers could determine the degree to which traits were genetically determined and test their effects on fitness, but how selection acted at the genetic level was unknown. The availability of large dense genetic marker sets has propelled this area of research forward along with new statistical methods of linking genotype to phenotype even in the face of complex population structure when using natural accessions (Zhao *et al.*, 2007; Kang *et al.*, 2008; Atwell *et al.*, 2010). One of these methods, genome-wide association studies (GWAS), has become commonly used to identify the genetic basis of quantitative variation in plants (Hall *et al.*, 2010).

Seed oil composition, the types and relative proportions of fatty acids found in seed oils, is a complex quantitative trait and the focus and all 3 chapters of my dissertation. Great variation in seed oil composition, both within and between species is maintained in nature (Sharma and Chauhan, 2012). Multiple studies have shown a

consistent pattern where plants grown in colder climates have a lower proportion of saturated fatty acids than those grown in warmer climates. Across over 680 species of angiosperms, those from tropical locations had a higher proportion of saturated fatty acids in their seed oils than in temperate species (Linder, 2000). Species within the genus *Helianthus* (Linder, 2000) and accessions of *A. thaliana* (Sanyal and Linder, 2013) followed a latitudinal cline where proportions of saturated fatty acids increase with decreasing latitude. The oilseed species *Plukenetia volubilis* L . displayed an inverse relationship between altitude and proportions of saturated fatty acids, where plants collected at higher altitudes (i.e., consistently colder temperatures) had significantly lower proportions of saturated fatty acids than plants collected from lower altitudes (Cai *et al.*, 2012).

Saturated fatty acids provide more energy per carbon atom (Lehninger, 1993) than unsaturated fatty acids of the same chain length and therefore one might expect all plants to maximize energy (if seed size and content were kept constant) by filling seeds with saturated fatty acids. However, saturated fatty acids have substantially higher melting points than unsaturated fatty acids when chain length is held constant (Eckey, 1954). Subsequently, seeds with low proportions of saturated fatty acids will also have a lower melting point causing them to be less viscous at cold temperatures. This characteristic forms the basis of an adaptive hypothesis that low melting point seeds (low proportion of saturated fatty acids) should be favored under colder germination temperatures due to

earlier germination and faster growth before photosynthesis, while at warmer germination temperatures seeds with a higher amount of energy (high proportion of saturated fatty acids) should be favored (Linder, 2000). Chapters 2 and 3 center on this adaptive hypothesis, with Chapter 2 devoted to testing the predicted emergence timing and fitness consequences under different germination temperatures. Chapter 3 uses natural accessions of *A. thaliana* to understand the genetic basis of adaptive quantitative variation in seed oil melting point.

While it has been shown that lower melting point seeds germinate earlier than seeds with higher melting points under cold germination temperatures (Miquel and Browse, 1994; Linder, 2000), relative fitness has not been assessed. The goal of Chapter 2 was to show whether seed oil melting point had an effect on emergence timing at different temperatures and whether that led to differences in relative fitness under competition. High and low melting point lines from a recombinant inbred cross of *A. thaliana* were competed in a fully factorial experiment at warm and cold temperatures with two different density treatments. Under warm temperatures, the predictions of the hypothesis of Linder (2000) were supported. There were no significant differences in emergence timing between high and low melting point lines, and plants under competition with high melting point lines had lower relative fitness than plants competing against low melting point lines. At cold temperatures, however, the lack of difference in emergence timing was unexpected and led to the same qualitative pattern of relative

fitness as under the warm temperatures. Plausible reasons for this deviation from expectation are discussed along with suggestions for a new experiment.

Previous work in *A. thaliana* has shown extensive variation in the relative proportions of the 9 fatty acids of the seed oils (Millar and Kunst, 1999; O'Neill et al., 2003; Hobbs et al., 2004; Sanyal and Linder, 2011) and a significant latitudinal cline in proportions of saturated fatty acids across accessions (Sanyal and Linder, 2013). Because more is known about fatty acid synthesis and triacylglycerol accumulation in *A. thaliana* than any other species (Wallis and Browse, 2010; Li-Beisson *et al.*, 2010; Li-Beisson *et al.*, 2013) and because of the availability of genome-wide SNP datasets and a well-annotated genome, *A. thaliana* was chosen as the study system to identify the genetic basis of variation in seed oil melting point (Chapter 3) and composition (Chapter 4).

Only 1 other study has attempted to uncover the genetic basis of the adaptive evolution of seed oil melting point in *A. thaliana* (Sanyal and Linder, 2011) but was limited to variation segregating between 4 sets of recombinant inbred lines. This study was able to identify 9 quantitative trait loci (QTLs) for seed oil melting point but they covered large genomic regions, each containing dozens to hundreds of genes. The work of Chapter 3 was devoted to increasing the resolution and scope of variation explored by using GWA to search for associations between the seed oil melting point variation of 391 *A. thaliana* accessions and over 214,000 genome-wide SNPs.

Genome-wide association mapping can also be used to identify the genetic basis

of quantitative traits in agriculturally important plant species, which can facilitate marker-assisted breeding for crop improvement or bioengineering (Mauricio, 2001; Myles et al., 2009; Rafalski, 2010; Al-Maskri et al., 2012). Particular seed oil compositions are very important in a number of practical uses of seed oils, e.g., highly reduced carbon for biodiesel, human consumption, and petrochemical replacement (Durrett *et al.*, 2008; Chapman and Ohlrogge, 2012; Vanhercke *et al.*, 2013). Many of the genes involved in the synthesis of fatty acids and their incorporation into triacylglycerols are known (Li-Beisson *et al.*, 2010; Li-Beisson *et al.*, 2013) but the regulation (Cernac and Andre, 2006; Baud *et al.*, 2007; Mu *et al.*, 2008; Baud and Lepiniec, 2009; Peng and Weselake, 2011) and alternative pathways of this system (Ståhl *et al.*, 2004; Zhang *et al.*, 2009; Xu *et al.*, 2012; Bates *et al.*, 2012) are still being uncovered and are presenting a picture of higher complexity than previously realized (Chapman and Ohlrogge, 2012; Vanhercke *et al.*, 2013). In chapter 4, I performed a GWA study of the relative proportions of the 9 fatty acids found in *A. thaliana* seeds oils and 4 composite fatty acid traits using 391 accessions of *A. thaliana* and a genome-wide SNP dataset. Pinpointing the genes underlying variation in seed oil composition in a world-wide set of wild *A. thaliana* accessions provides a set of target genes for manipulation of seed fatty acids proportions in agriculturally important oilseeds due to conservation of structure and function of these genes in many plant species (Sharma and Chauhan, 2012).

Chapter 2. Emergence timing and fitness consequences of variation in seed oil melting point

Introduction

Timing of germination and seedling emergence can contribute significantly to the overall lifetime fitness of plants, particularly annuals. Many studies have found evidence of a selective advantage to early emergence, especially under competition (Miller *et al.*, 1994; Dyer *et al.*, 2000; Geber and Griffen, 2003; Mercer *et al.*, 2011). Seed oil composition (the types and relative amounts of fatty acids in the oils) may play an important role in determining germination timing and emergence for the 80% of flowering plants that store oils for energy in their seeds (Linder, 2000). These seed oils provide the energy to fuel germination and emergence prior to the onset of photosynthesis (Harwood, 1980). Consistent patterns in the composition of seed oils have been found across the angiosperms along a latitudinal gradient at both the family and genus levels and within broadly distributed species, which is indicative of selection acting on the relative proportions of fatty acids (Linder, 2000). Across 680 oilseed species, tropical species had significantly higher proportions of saturated fatty acids in their oils on average than temperate species. Oil composition within the wide-spread genus *Helianthus* was significantly negatively correlated with latitude (Linder, 2000).

This micro- and macro-evolutionary evidence suggests that selection has acted to repeatedly modify saturation level of the fatty acids in seeds oils in response to

germination temperatures (Linder, 2000). An adaptive hypothesis to explain this pattern suggests that under competition, selection has acted to modify seed oil composition through a trade off between total amount of energy stores in seeds and the rate at which that energy can be used at different germination temperatures (Linder, 2000). Saturated fatty acids are less costly to produce and provide more energy per carbon atom (Lehninger, 1993) but have much higher melting points than unsaturated fatty acids of the same chain length (Eckey, 1954; Williams, 1964). Therefore, oilseeds with higher proportions of unsaturated fatty acids should have lower melting points and should be able to germinate earlier and grow more rapidly under colder germination temperatures than competitors with a lower proportion of unsaturated fatty acids in their seed oils. Under warmer germination temperatures, this timing advantage should disappear, and the higher amount of total energy, in the form of a higher proportion of saturated fatty acids, should be favored. Using *H. annuus* seed having high or low amounts of saturated fatty acids in their seed oils, there was not a significant difference in germination timing at a warm germination temperature, while at lower germination temperatures, the seeds with low amounts of saturated fatty acids germinated significantly earlier (Linder, 2000) and grew more rapidly. This experiment did not test for the expected differences in fitness.

With short generation times and small stature, the weedy annual *Arabidopsis thaliana* is well suited to studies of life history traits in large-scale growth chamber experiments. Natural populations of *A. thaliana* experience a broad range of

temperatures and other climatic conditions across the species' native Eurasian range (Lasky *et al.*, 2012) providing the opportunity for selection on life history characters within different populations. Previous studies have found genetic variation in *A. thaliana* for life history traits, e.g., seed dormancy flowering time (Debieu *et al.*, 2013), and have shown that this variation can affect lifetime fitness (Huang *et al.*, 2010; Donohue *et al.*, 2005). The proportions of saturated fatty acids in A. thaliana seed oils have been shown to follow a latitudinal cline, with an inverse relationship (Sanyal and Linder, 2013). This dataset was used to estimate seed oil melting points, which were more strongly correlated with latitude that saturated fatty acid proportions (data not shown). Because of this strong relationship, the lines for this study were chosen based upon seed oil melting point differences.

In this study, I investigated the impact of temperature on time to first emergence in 4 *A. thaliana* recombinant inbred lines (RILs) with disparate seed oil melting points (two with high melting points and two with low) and measured relative fitness. Under the hypothesis of Linder (2000), I had two expectations. First, under lower germination temperatures, lines with lower melting points would have higher relative fitness under competition with a higher melting point line due to earlier emergence than high melting point lines under competition with lower melting point lines. Second, under warm germination temperatures, I expected no difference in emergence timing between high and low melting point lines and for the higher melting point lines to either have similar or

higher fitness relative the lower melting point lines when in competition with individuals of the opposite type. I tested these hypotheses using fully factorial intra- (high v high or low v low) and inter-type (high v low or low v high) competition experiments under controlled temperature conditions using fruit count as the proxy for fitness.

Materials and methods

Plant materials and seed production

I used four lines from a previously described set of 100 recombinant inbred lines (RIL set CS1899), generated from a cross between *A. thaliana* accessions Columbia (CS933) and Landsberg erecta (CS20) (Lister and Dean, 1993). Two lines were chosen from the upper and two from the lower tails of the distribution of seed oil melting points of this RIL set. Melting points were estimated from the RIL set's published seed oil compositions (Sanyal and Randal Linder, 2011) by calculating an average of the melting points of the fatty acids, weighted by their relative proportions (Malkin, 1954). The low melting point lines (L) had melting points of 9.9ºC (CS1975) and 10.0ºC (CS1984), while the high melting point lines (H) had melting points of 14.5ºC (CS1948) and 14.6ºC (CS1994).

Seeds for these lines were purchased from the Arabidopsis Biological Resource Center (ABRC). To break dormancy, seeds were surface-sterilized and cold stratified in water for ten days at 4ºC prior to planting. Plants were grown under common garden conditions of 15 hrs light at 22ºC and 9 hours of dark at 18ºC. Moist soil was maintained through subirrigation and plants were fertilized once a week with Peters professional 20-

20-20 (Everris NA, Inc.). Upon bolting, plants were kept reproductively isolated using the Arasystem (Betatech, Gent, BE). Seeds were harvested from each plant after the siliques matured and turned brown. The seed produced from these plants was stored at room temperature in unsealed coin envelopes for 20 months prior to the start of this experiment.

Experimental design

A fully factorial randomized-block design was implemented for the competition trials under two germination temperatures (Cold: 11ºC day/6.5ºC night, and Warm: 16ºC day/11ºC night) and at two densities (High: 1cm between the focal and associates, and Low: 3 cm). Cold temperatures were chosen so that low melting point lines would be liquid/low viscosity during the day but the high melting point lines would not. Warm temperatures were higher than the melting point of all 4 lines to prevent significant seed oil viscosity differences between the 4 lines. Following stratification, seeds were sown into either 2.5" diameter pots (high density treatment) or 3" pots (low density treatment). Pots were filled with Metro-Mix 360 (SunGro Horticulture Canada Ltd., Vancouver, Canada). Competition pots consisted of a focal plant surrounded by 4 competitors of one genotype (high or low melting point). The focal was sown into the middle of the pot and the competitors were added to the four corners of the pots at distances of 3cm from the focal in the low density treatment and 1cm in the high density treatment. All possible combinations of focal genotype with competitor genotype were planted (4 lines x 4

lines=16 combinations). No-competition controls (focal plant only) for each of the 4 lines were planted in both pot sizes to test whether the density treatments induced competition. Seeds were sown onto the surface of 8 replicate pots for each factorial combination of each planting combination for a total of 640 pots (2 densities x 2 temperatures x 20 planting combinations x 8 replicates). Pot position was randomized across 5 trays of 2.5" pots and 8 trays of 3" pots in two growth chambers for a total of 26 trays (13 per growth chamber). Tray was treated as a block and tray positions were randomized twice per week to minimize position effects.

The temperature treatments were applied during the germination and establishment phase in two different growth chambers. The light intensity was equivalent in both chambers and lights were on 14 hours per day.

All measurements were recorded for only focal plants. Pots were observed daily and date of first emergence recorded when cotyledons appeared. After no new focals had emerged over a five day period, plants were moved to a single growth chamber to finish their life cycle under common garden conditions of 16 hours of light at 23ºC and 8 hours of dark at 18ºC. Trays were subirrigitated at a frequency that maintained moist soil in pots and were fertilized once per week with Peters professional 20-20-20 (Everris NA, Inc.). A 9-week growing season was enforced by discontinuing watering after that time and removing newly flowering meristems. To assess relative fitness, total fruits per focal plant were counted. Plants that died before reproduction were assigned a fruit count of 0.

Statistical analyses

Because the adaptive hypothesis of Linder (2000) makes predictions about emergence timing and relative fitness specific to each temperature treatment, data from the two temperatures were analyzed with separate models. The effects on days to emergence of focal type (H or L), density, and their interaction (all incorporated as fixed effects) were assessed using generalized linear mixed models (proc GLIMMIX in SAS 9.2). The best fit distribution for emergence days in the cold model was gaussian and poisson for the warm model.

To analyze the effects of focal/competitor type, density, and their interaction (all incorporated as fixed effects) on fruit count, proc GLIMMIX was again used but with a negative binomial distribution for both cold and warm models. Orthogonal contrasts were added to the fruit count models to test the hypotheses of Linder (2000). First, the relative fitness consequences of seed oil melting point predicted by Linder (2000) apply to plants under competition, so contrasts between competitor arrays of a particular focal type (H or L) and no-competition controls of the same focal type were performed to assess whether competition was induced at the given planting densities.

Contrasts of intra- and inter-type fruit counts for a given focal type (ex. LH v LL) were used to address the specific relative fitness predictions of Linder (2000) in both the warm and cold temperature models. Under cold temperatures, the adaptive hypothesis (Linder, 2000) predicts higher relative fitness for high melting point lines under intra-type

(HH) competition compared to inter-type (HL), and higher relative fitness for low melting point lines under inter-type competition (LH) compared to intra-type (LL). Under warm temperatures, focal plants (regardless of type) are predicted to have higher fitness when in competition with low melting point lines relative to those in competition with high melting point lines.

Results

Time to first emergence

Temperature had a strong effect on time to first emergence, with an average time of 7.9±0.036 days in the cold treatment and approximately half the time in the warm treatment (mean±SE = 4.1±0.026) (Figure 2.1). As expected density had no detectable effect on emergence time, but neither did melting point type (H or L). Within a given temperature treatment, no emergence time differences were significant (Table 2.1). However, low melting point type focal plants emerged on average 0.10 days earlier than high melting point lines under cold temperatures, which was a non-significant trend in the predicted direction (p value=0.15). Which factors were significant did not change whether tray was included as a random variable or excluded from any of the models for either emergence days or fruit count, so the results excluding tray are presented.

Density treatments induce competition

In all cases, at both planting densities, the no competition plants produced significantly more fruits (567.7±29.1) than the plants with competitors (155.5±5.1)

demonstrating that I was successful at creating competitive conditions for plants with associates.

High density plants, whether competition or no-competition, grew in smaller pots than the low density plants. Pot size had a significant effect on fruit number for high- and low-density no competition plants with a mean difference of 389 more fruits per plant for plants growing at low density (Figure 2.2). Within each temperature treatment, density had a highly significant (p value<0.0001) effect on fruit number overall and within each focal type between controls and those under competition (ex. LH and LL v LN) (Table 2.2 and 2.3). On average, plants with competitors at 1 cm distance produced 3.9 times more fruits than the no-competition controls and 3.5 times more at the 3 cm distance.

Relative fitness

Cold germination temperatures

Under cold germination temperatures, the first expectation was that lines with low melting point oils would have a higher relative fitness when in inter-type (LH) competition than intra-type (LL) competition and this relationship would be stronger under higher densities. This was not observed, as there was not a significant difference in fruit number produced by the low melting point lines under the two competition types and there was not a significant interaction between focal/competitor type and density (Table 2.2). Contrasts between HL and HH, while not significant, trended in the opposite direction of expectation with inter-type competition having higher fitness (Figure 2.2).

The second prediction was that high melting point lines would have higher fitness under intra-type (H v H) competition relative to inter-type (H v L). The opposite pattern was significant (p value=0.014, Table 2.2) with an inter-type mean fruit number of 210.3±35.2 and an intra-type mean of 160.0±31.0 (Figure 2.2).

Warm germination temperatures

Under warm germination temperatures, the adaptive hypothesis predicted higher intra-type fitness for low melting point lines relative to inter-type competition. Although non-significant (p value=0.19, Table 2.3), the mean fruit number (138.9±23.8) of the intra-type set (LL) was higher than the inter-type set (115.0±17.1).

As expected, high melting point lines had significantly higher reproductive success (p value=0.012) when in competition with low melting point lines (217.3±36.1) than with the intra-type (mean fruit count=152.5±24.4) (Figure 2.2).

Discussion

My goal was to determine the effect of seed oil melting point on timing to emergence under different germination temperatures and the resulting consequences, if any, on relative fitness. The only significant difference in timing to first emergence was between temperature treatments, with first emergence of seeds in the warm treatment in half the time of the cooler temperature. No difference in emergence timing was expected between seeds with different oil melting points under warm temperatures, but the lack of timing differences was unexpected under cold germination temperatures. There was,

however, a non-significant trend of seeds with lower melting points emerging earlier than those with higher melting points, following the direction of predictions under the hypothesis of Linder (2000).

The adaptive hypothesis of Linder (2000) for the production of seed oil composition variation in nature hinges upon differences in germination timing between high and low melting point lines grown under cold temperatures. In the case of *A. thaliana,* emergence timing is a reasonable proxy for germination timing because the seeds are on the soil surface and emergence of the cotyledons quickly follows germination. There are at least three possible explanations for the lack of significant emergence timing differences of the high and low melting point RIL lines in this study.

First, the hypothesis that the latitudinal cline of melting points is driven by selection on germination timing could be incorrect. Two direct experiments have investigated the effect of temperature on germination timing of seeds with different melting points. In the first, the germination timing of high oleate fad2-2 mutant seeds were compared with wild type *A. thaliana* seeds under 3 different germination temperatures (Miquel and Browse, 1994). At 6ºC and 10ºC, the wild-type seeds (lower melting point) germinated significantly earlier than the mutant fad2-2 seeds (higher melting point), while at 22ºC germination timing differences were non-significant. In the second experiment, *H. annuus* seeds with lower proportions of saturated fatty acids germinated significantly earlier than those with higher proportions at cooler temperatures,

while the difference of timing was indistinguishable at warmer temperatures (Linder, 2000). Both of these studies confirm the prediction of Linder (2000) that lower melting point lines will germinate earlier under cold temperatures. This prediction is also supported by observational evidence from nature. Across the angiosperms, tropical species on average have a higher proportion of saturated fatty acids than temperate species (Linder, 2000). Lower latitude species (consistently warmer temperatures across the Midwest) in the genus Helianthus have a significantly higher proportion of saturated fatty acids (Linder, 2000). This same latitudinal cline of proportions of saturated fatty acids increasing with decreasing latitude has been observed in *A. thaliana* (Sanyal and Linder, 2013). Plants from the oilseed species *Plukenetia volubilis* L . collected at higher altitudes had significantly higher proportions of unsaturated fatty acids than plants collected from lower altitudes (ie. consistently warmer temperatures (Cai *et al.*, 2012)). In the face of the evidence in support of this adaptive explanation, further testing with appropriate changes in experimental design seems worthwhile.

Second, the non-significant trend of earlier emergence of low melting point might be due to insufficient power to reach significance. If that is the case, the relationship is a weak one and one that in this experiment did not produce the expected fitness consequences (see below). Only 2 lines from either end of the melting point distribution of the RIL set were chosen, leading to qualitative melting point factors in the analysis (H v L). More power to detect consistent differences in emergence timing would be provided

by using more lines so that a quantitative approach to melting point could be incorporated in the model.

Finally, the difference between the melting points of the high and low types may have been too small to have a large effect on germination timing under the temperatures used for this experiment. The cold germination temperature may not have been low enough to drive viscosity changes needed for emergence timing differences between high and low melting point lines. While here I used 11ºC daytime/6.5ºC nightime temperatures for the cold treatment, both of the previous experiments used lower cold temperatures, 10ºC daytime and 4ºC nighttime for Linder (2000) and 10ºC and 6ºC for Miquel (1994). The range in melting point between the low melting point lines and the high melting point lines was from 9.9ºC to 14.6ºC, for a 4.7ºC difference. Wild accessions of *A. thaliana* have been exposed to natural selection pressures and have a wider range of seed oil melting points (8.7ºC to 15.8ºC) with a difference of 7.1ºC and therefore would make a more appropriate study system.

Because a significant difference in emergence time was not observed under either temperature treatment, the greater proportions of high energy saturated fatty acids found in the higher melting point lines might have given them a competitive advantage. Seed oils provide the energy to germinate and grow until at least the onset of photosynthesis and therefore, maximizing this energy source should be beneficial. Because saturated fatty acids provide more energy per carbon atom than unsaturated and are cheaper for the

mother plant to produce (Lehninger, 1993), it follows that with a given seed size and oil content, more energy can be provided to the growing seedling by filling it with saturated fatty acids. In support of this idea, within every treatment, plants of both types competing against high melting point lines (14.2% and 14.6% saturated fatty acids) had lower relative fitness than those competing against lower melting point lines (11.5% and 12.1% saturated fatty acids), although this relationship was not significant with low melting point focal plants (ie. LL > LH).

If this experiment were redone with colder germination temperatures and low melting point lines emerged significantly earlier than high melting point lines the fitness differences predicted by Linder (2000) would be expected. The predictions would be that low melting point lines would have higher inter-type (LvH) fitness relative to intra-type competition (LvL) and high melting point lines would have higher intra-type (HvH) fitness relative to inter-type competition (HvL).

Chapter 3. Adaptive genetics of seed oil melting point

Abstract

Seed oil melting point is an adaptive quantitative trait determined by the relative proportions of the fatty acids that compose the oil. Micro- and macro-evolutionary evidence suggests selection has changed the melting point of seed oils to covary with germination temperatures because of a trade-off between total energy stores and the rate of energy acquisition during germination under competition. In this study, the seed oil compositions of 391 natural accessions of *Arabidopsis thaliana* grown under common-garden conditions were used to assess whether seed oil melting point varied with germination temperature and to determine which genes were most likely responsible for the variation in seed oil melting point. In support of the adaptive explanation, long-term monthly spring field temperatures of the accession collection sites significantly predicted their seed oil melting points, particularly temperatures in February. Using methods that control for population structure, a genome-wide association (GWA) study of seed oil melting point discovered 12 genes tightly linked with SNPs significantly associated with seed oil melting point variation. Seven of these genes encode either lipid synthesis enzymes or trans-acting regulatory products, while the remaining 5 genes encode products with no clear relation to seed oil melting point. The seed oil compositions of knockout lines for the two *a priori* candidate genes found to be significantly associated with seed oil melting point, *FAD2* and *FATB*, affected seed oil melting point as expected

with the *FAD2* line increasing melting point by 9.6ºC relative to wild-type and the *FATB* knockout causing a 3ºC decrease. Using methods that did not control for population structure, I identified several additional genes that encode products with a clear mechanism for alteration of seed oil melting point. Our results suggest selection can act to alter quantitative trait variation in response to local environmental conditions through a small set of genes.

Introduction

Understanding the genetic basis of adaptive evolution of quantitative traits, one of the major areas of evolutionary research, has become attainable. Recent genotyping advances and methods for utilizing large genetic data sets have led to an increase in research addressing this goal (Brachi *et al.*, 2010; Childs *et al.*, 2010; Chan *et al.*, 2011; Hancock *et al.*, 2011). Evolution requires the presence of heritable genetic variation in traits and most traits are quantitative. Therefore to understand adaptive evolution, I must understand the genetic architecture of complex trait variation. This knowledge also has practical applications, such as identifying genes underlying human disease (Stein and Elston, 2009) and those important for crop improvement (Rafalski, 2010).

Evidence supports adaptive evolution of seed oil composition, the types and relative proportions of the fatty acids in seed oil, in angiosperms that produce oil seeds (Linder, 2000). This oil supplies the energy for germination and establishment prior to photosynthesis (Rees and Long, 1992). Seed oil composition determines both the total

amount of energy stored in the seed (when oil content is held constant) as well as the melting point of the oil, which affects the rate at which the energy can be used at different temperatures. Seed oil melting point decreases with the degree of fatty acid desaturation and with decreasing chain length, making desaturated and shorter fatty acids better suited to germination at lower temperatures. Per unit of change, desaturation has a stronger effect on melting point than decreasing chain length. On the other hand, saturated fatty acids cost less to synthesize and provide more energy per carbon atom when metabolized than unsaturated fatty acids. Therefore, increased relative amounts of unsaturated fatty acids decrease seed oil melting point at the expense of total energy stores. In locations with cooler germination temperatures, seeds with higher proportions of unsaturated fatty acids are expected be able to germinate earlier and grow more rapidly prior to photosynthesis, which would be advantageous under competition. Under warmer germination temperatures, this advantage disappears and seeds with a higher amount of total energy stores, i.e., proportionately more saturated fatty acids, may be favored. Therefore, the adaptive evolution of seed oil composition is thought to be a trade-off between total energy stores and the rate those stores can be used, with germination temperature as the selective agent (Linder, 2000).

Based on this adaptive explanation, genes encoding enzymes that desaturate fatty acids would be expected to play an important role in modulating seed oil melting point. Most of the enzymes involved in seed oil biosynthesis are well understood and their

functions validated in mutant and molecular studies (reviewed in Li-Beisson et al., 2010). While the enzymes involved in the biosynthesis of seed lipids have been identified, relatively little is known about how lipid synthesis is regulated to produce variation in fatty acid composition (Li-Beisson *et al.*, 2010). In addition to the genes in the oil synthesis pathway, genes encoding trans-acting regulatory factors such as kinases and transcription factors could act to control proportions of fatty acids produced and, therefore, be important in modifying the seed oil melting point (Ruuska *et al.*, 2002).

The adaptive evolution of seed oil composition, seen at the family and genus levels (Linder, 2000), would also be expected within a species with a broad geographic distribution. Previous work has shown extensive variation in seed oil composition in natural accessions of *A. thaliana* (Millar and Kunst, 1999; O'Neill *et al.*, 2003), and this variation often follows the same pattern as seen among species (Sanyal and Linder, 2011).

To date, the study of the quantitative genetics of seed oil composition has been confined to quantitative trait loci (QTL) studies (Hobbs *et al.*, 2004; O'Neill *et al.*, 2011), and only one of these has examined the quantitative genetics of the melting point of seed oil (Sanyal and Linder, 2011). In the Sanyal and Linder study, the QTL peaks covered large genomic regions spanning dozens to hundreds of genes.

With a world-wide distribution and the availability of abundant genetic resources —particularly dense SNP data sets—*A. thaliana* is an excellent system to determine the

fine-scale genetic basis of adaptive evolution. If enough markers are used, associations can be resolved to the genic level in most instances (Bergelson and Roux, 2010; Weigel, 2012). GWA studies are now possible in species with a high degree of population structure, like *A. thaliana,* because of new statistical tests that look for associations while controlling for relatedness (Kang *et al.*, 2008). Controlling for relatedness reduces the false positive rate because some markers can be confounded with the population structure. On the other hand controlling for population structure also can lead to an increase in false negatives especially for traits that have been selected along the same geographic lines as the population structure (Brachi *et al.*, 2010). By using multiple types of association mapping (QTL, and GWA with and without controlling for population structure), the genetic variation contributing to trait variation in nature can be identified while minimizing false positives and negatives (Brachi *et al.*, 2010; Bergelson and Roux, 2010). There are several examples of this in the literature (e.g., Brachi et al., 2010; Li et al., 2010; Chan et al., 2010; Nemri et al., 2010).

The goal of this study was to identify the genetic basis of the observed variation in the seed oil melting point of *A. thaliana* and to assess whether the genes identified shed any light on the adaptive evolution of seed oil melting point. I were particularly interested in assessing whether any genes from a list of 15 *a priori* candidate genes for melting point (Sanyal and Linder, 2011) would be implicated, and, if not, what new candidate genes would be identified. Using the fatty acid compositions and estimated seed oil

melting points from 391 natural accessions of *A. thaliana*, a study was performed on the *a priori* candidate genes and a GWA study was performed to identify *a posteriori* candidate genes. Tests were performed that did and did not control for population structure. The results are compared to the QTL mapping results of Sanyal and Linder (Sanyal and Linder, 2011).

Materials and methods

Accessions and seed production

Four hundred forty-seven putative spring-germinating accessions from across the geographic range of *A. thaliana* were chosen to maximize representation of natural variation in seed oil composition. Most were within the native range of *A. thaliana* (392 European, 6 Asian, 12 from the border between Europe and Asia, and 1 African), and 36 accessions were recent introductions outside of its native range (United States 29, Japan 3, Cape Verde Islands 1, Canada 2, and New Zealand 1). Seeds were acquired through the Arabidopsis Biological Resource Center (ABRC) and the Juenger lab at the University of Texas at Austin.

Surface-sterilized seeds were imbibed and stratified at 4ºC for 6 days to break dormancy prior to planting. Following stratification, plants were grown from November 2010 to March 2011 in a greenhouse with 16 hours of supplemental lighting from high pressure sodium high-intensity discharge lights. Growing temperature ranged from 6.1°C to 28.9°C (mean=18.1°C). Three replicates of each accession were planted in separate

pots in a randomized block design for a total of 1,341 pots split across 77 trays. An average of five seeds were planted in each pot and randomly thinned to one plant after germination and establishment. To decrease environmental variation, especially due to edge effects, pots were placed in every other tray slot such that every pot was equidistant. Twice per week the trays were subirrigated, fertilized with Dyna-Gro 7-9-5 plant food (Dyna-Gro Nutrition Solutions, Richmond, CA), and tray locations within the growing area were randomized. Once a plant bolted, it was kept isolated using the Arasystem (Betatech, Gent, BE). Seeds were harvested from each plant after the siliques matured and browned. Seeds were stored in unsealed coin envelopes at room temperature for 2 to 9 months before seed oil compositions were determined.

Seed oil composition analysis

Fatty acid compositions of the accessions were determined using gas chromatography of fatty acid methyl esters (FAMEs). Extraction procedures were modified from (Zheljazkov *et al.*, 2008): about thirty randomly chosen seeds per plant were crushed with a glass pestle in a 2mL glass autosampler vial and mixed with 150 L of extraction buffer (75% hexane, 20% chloroform, and 5% 0.5 M sodium methoxide in methanol). A minimum of five minutes elapsed before the first samples were injected into the chromatograph. Two extractions were performed per plant, for a total of six samples per accession. Samples were analyzed using an HP 5890A gas chromatograph with a robotic autoloader and a DB-23 capillary column (Agilent Technologies, Santa Clara,

CA). The initial oven temperature of 180°C was maintained for 5.5 min and then raised 15°C per min to a final temperature of 240°C for 0.5 min for a total run time of 10 min. Two injections were performed for every tenth sample to ensure the repeatability of the results. FAME peaks were visualized using a flame ionization detector and were identified by comparison to FAME standards 189-4, 189-5, 189-12 and 189-19 (Supelco, Bellefonte, PA). The standards were run before each set of seed oil extractions. Peaks were integrated using Agilent Chemstation software revision A.04.02, and the relative proportions of the nine principle fatty acids in *A. thaliana* seed oil (16:0, 18:0, 18:1, 18:2, 18:3, 20:0, 20:1, 20:2, and 22:1) were measured. The first number corresponds to the number of carbon atoms in the fatty acid chain and the second the number of double bonds. For each sample, the proportions of fatty acids were used to calculate three integrated traits that might be biologically important based on knowledge of fatty acid biochemical pathways (http://aralip.plantbiology.msu.edu). The trait 'Plastid' was the sum of the proportions of the three fatty acids (16:0, 18:0, and 18:1) produced in the plastid. Because saturated fatty acids proportionately increase the overall melting point of seed oils more than unsaturated fatty acids, the total proportion of saturated fatty acids (Sat) was calculated by adding the proportions of 16:0, 18:0, and 20:0. Finally, seed oil melting point (MP) is putatively adaptive (Linder, 2000) so it was estimated as the average of the melting points of the individual fatty acids weighted by their relative proportions (Malkin, 1954).

Relative effect of fatty acids on seed oil melting point

The effect of each fatty acid on the seed oil melting point is dependent not only on the relative proportion of that fatty acid but also the degree to which its individual melting point differs from the overall seed oil melting point. The effect of each fatty acid was calculated as follows: effect = [(fatty acid MP - seed MP)/ seed MP] * fatty acid proportion. This calculated value will be referred to as the melting point effect metric.

Climate and geographic regressions

Using patterns of isolation by distance, the geographic origins of 63 of the accessions used in this study were previously reported as likely misidentified (Anastasio *et al.*, 2011). Therefore, for the following climate and geographic linear regressions, only the remaining 328 accessions were used. To test whether germination temperatures were predictive of seed oil melting point, as would be expected if the melting point is adaptive, linear regressions were performed in R using the minimum, mean, and maximum temperatures for the months of February through May as estimated by (Hijmans *et al.*, 2005). The monthly temperature data specific to the accessions in this report was obtained from a previous study (Lasky *et al.*, 2012).

To determine whether MP varied along the same geographic lines as population structure (Weigel, 2012), linear regressions were performed in R using latitude, longitude and altitude data for the accessions.

SNPs

214,051 SNPs were downloaded from http://walnut.usc.edu/2010/SNPs (v3.06). The method for obtaining these SNPs was previously described in (Atwell *et al.*, 2010). Of these, I removed 42,603 SNPs with allele frequencies of 10% or less from the dataset to control for the effect of minor alleles on the analysis (Kang *et al.*, 2008; Atwell *et al.*, 2010).

Association mapping

Of the 447 accessions planted, 8 did not germinate, 2 died before producing seed, and 13 did not have genotypic information available in the SNP dataset. A further 33 lines required a vernalization period to initiate flowering, indicating they were incorrectly typed fall-germinants. None of these accessions were used for the association mapping study. The remaining 391 accessions were used in all analyses (Figure 3.1). Using multiple measurements per accession can increase the power to detect associations, so all 6 replicates were used independently in the statistical analyses (Kang *et al.*, 2008).

Genome-wide association mapping was performed for seed oil melting point. A total of four analyses were conducted using efficient mixed model association (EMMA) (Kang *et al.*, 2008): with and without controlling for population structure and with and without non-native accessions. I used the EMMA.reml.t function of the R implementation EMMA v. 1.1.2 for Linux. EMMA identifies associations between SNP genotype and trait variation while controlling for genetic relatedness using a pairwise genetic similarity

matrix (Kang *et al.*, 2008). The SNP genotype is modeled as a fixed effect and the genetic similarity matrix as a random effect. Although controlling for population structure reduces the false positive rate, it can also lead to increased false negatives when population structure is confounded with selection on a trait. Because seed oil composition varies along some of the same geographic clines as the population structure of these accessions (Weigel, 2012), false negatives were expected from the EMMA output. I therefore performed a second association mapping study with all values in the kinship matrix set to zero (referred to as the no-population-structure analysis).

Since 36 of the accessions in this study were from outside of the native range of *A. thaliana*, they may have arrived at their current locations too recently for selection to have acted on their seed oil melting point or they may have gone through a genetic bottleneck (founder effect) that limited their ability to respond to selection. I therefore conducted both types of EMMA analyses on the full set of accessions (N = 391) and on only the native accessions (N = 356).

Because association mapping using such a large dataset is computationally expensive and EMMA models each SNP independently, the dataset was divided into 1,020 sets of 210 SNPs each and run in parallel on the Lonestar cluster of the Texas Advanced Computing Center at the University of Texas at Austin.

Benjamini-Hochberg FDR corrections (Benjamini and Hochberg, 1995) were done on all p-values output from EMMA to control for multiple testing. An FDR

threshold of 0.05 was chosen to identify significantly associated SNPs.

A priori candidate genes

A set of 15 *a priori* candidate genes known to encode fatty acid synthesis enzymes was chosen as likely to affect seed oil melting point (Sanyal and Linder, 2011). Six of these genes collocated with QTL for seed oil composition. A separate FDR correction was applied to the p-values of SNPs that fell within 1000 bp upstream or downstream of the coding regions of the *a priori* candidate genes and examined separately from the genome-wide analysis.

Identification of significant genes

To identify the most likely causal variants linked to the significant SNPs, two criteria were followed. First, a gene or its promoter was considered likely to contain the causal variant if a significant SNP fell between its start and stop codons. However, due to the uneven distribution of SNPs across the genome, not all genes have SNPs within them. In addition, a gene with the causal variation could contain non-significant SNPs that have arisen since a selective sweep. Second, for SNPs not within gene boundaries, the closest gene was considered linked.

Gene functions and GO terms were obtained from TAIR (www.arabidopsis.org) and http://aralip.plantbiology.msu.edu/.

Results

Fatty acid composition and melting points

Seed oil melting point (MP) varied across the 391 accessions by 7.1°C, from 8.7°C to 15.8°C with a mean of 13.6°C (Figure 3.2). The distribution of melting points was mostly normal with several low melting point outliers. These values did not vary significantly between the full set of accessions used (All, N=391) and the accessions found in the native range of *A. thaliana* (Native, N=356). Of the four fatty acids with the greatest value for the seed oil MP effect metric (Figure 3.3), three fatty acids (18:2, 18:3, and 20:1) were present at the highest average proportions in the seed oil, (Figure 3.4), while the proportionately low 16:0 had a large effect because of its comparatively high melting point (62.9°C) relative to that of the seed oil (mean=13.6°C). 18:2 and 18:3 most strongly decreased the melting point, while 16:0 had the strongest increase on MP followed by 20:1 (Figure 3.3).

Climate and geographic regressions

The minimum, mean and maximum temperatures of a collection locale for each of the months of February through April significantly predicted seed oil MP (Table 3.1). Minimum temperature was the best predictor across all three months, with minimum temperature in February being the best overall ($r^2 = 0.169$; Figure 3.5). Mean temperatures were the next best predictors followed by maximum temperature. The

minimum February temperature had the strongest relationship with seed oil MP of any of the climate variables by an order of magnitude. The only geographic variable that significantly explained some of the variation in MP was longitude ($r^2 = 0.078$).

Genome-wide association study

SNPs associated with seed oil melting point

Manhattan plots revealed several areas of strong association with MP (Figure 3.6). Fourteen SNPs were significantly associated with seed oil MP in the tests with-population-structure (FDR<0.05). Ten were identified by both the All and Native sets, while three were found only by the Native test and one was unique to the All test. The main axis of population structure in *A. thaliana* has been shown to be longitudinal (Sharbel *et al.*, 2000; Nordborg *et al.*, 2005; K., J., Schmid *et al.*, 2005; Ostrowski *et al.*, 2006; Beck *et al.*, 2008; Picó *et al.*, 2008; Platt *et al.*, 2010; Cao *et al.*, 2011), and it was clear that longitude significantly predicted seed oil MP (Table 3.1), therefore the tests that control for population structure are expected to produce false negatives for any markers that vary along this gradient and our analyses that ignored population structure, as expected, identified a very large number of significant markers.

Even after FDR correction (FDR<0.05), the tests that did not control for population structure identified very large numbers of significant MP SNPs (All=34,657 and Native=32,981), suggesting MP very frequently was confounded with population structure producing a high proportion of false positives. To help reduce these significant

SNPs to those most likely associated with MP, only the 100 most significant SNPs from the All and Native sets were used in the rest of the analyses (Figure 3.6). Eighty-one of these SNPs were common to both tests and 19 unique SNPs were identified in each test, for a total of 119 significant MP-associated SNPs for the no-population-structure tests. Of the 14 significant MP SNPs identified by the with-population-structure tests, 8 were also significant in the no-population-structure tests.

One of the most visually striking signals in the genome-wide Manhattan plot (Figure 3.6), a cluster of significant SNPs on chromosome 5 spanning about 30 kb (Figure 3.7), was significantly associated with seed oil MP in both tests. This region had two distinct peaks of low p-values in both tests but only one was significant in the with-population-structure tests, suggesting the SNPs in the second peak could be false negatives.

Colocation of significant SNPs to previously identified QTL

While the MP-associated SNPs identified by both tests showed enrichment in the QTL regions previously identified by (Sanyal and Linder, 2011), enrichment was greater in tests with-population-structure, suggesting a successful reduction of false positives. The 95% CIs of these nine MP QTL covered approximately 26.89 Mb (19.97% of the genome). Of the 14 SNPs associated with MP in the tests with-population-structure, 11 were found within these QTL (Table 3.2), which is 3.9 times more than would be expected at random. In addition, 63 of the 119 significant MP SNPs identified by the tests

without-population-structure were in these QTL, 2.7 times more than expected by chance (Table 3.3). Three of the 9 QTL (chromosomes 1, 3, and 5) had significant SNPs in the no-population-structure tests but not in the tests with-population-structure, which suggested these SNPs might be false negatives.

Interestingly, a large proportion of the significant SNPs identified by both the with-population-structure (7/14) and no-population-structure (27/119) tests were found within the 95% confidence interval of the first QTL on chromosome 5, reinforcing the importance of this region on MP variation.

Genes linked to significant SNPs

While the desaturases of the *a priori* candidate genes were predicted to affect MP based on the adaptive hypothesis (Linder, 2000), other enzymes involved in lipid metabolism were also expected to be associated. Genes that encode products able to trans-regulate lipid synthesis were expected to affect MP variation as well. Several genes in these two categories were considered linked to SNPs found to be significantly associated with seed oil MP.

Significant SNPs were considered linked to the nearest gene. Nine of the 14 with-population-structure MP-associated SNPs (Table 3.2) and 98 of the 119 no-population-structure MP SNPs (Table 3.3) were located within genes. For the remaining SNPs, the closest gene ranged from 305 bp to 6006 bp away, which is within the 10 kb estimate of linkage disequilibrium decay in *A. thaliana* (Kim *et al.*, 2007). A total of 88 unique genes

were identified as MP-associated, 12 by the with-population-structure tests, 83 by the no-population-structure tests, with 7 overlapping.

Genes for Lipid Related Enzymes

Ten of the 88 MP-associated genes identified in this genome-wide study encode products with functions known to be related to lipid synthesis, degradation or transport. Of these ten genes only *NADP-ME2 (AT5G11670)* was identified by both tests, *FAD2 (AT3G12120)* was unique to the with-population-structure test, and 8 were identified only by the no-population-structure test. They will be reported from most significant to least significant within test type in the following sections.

With-population-structure analyses

FAD2 contained the two most significant MP SNPs in the test with-population-structure, with p-values an order of magnitude lower than any other SNP in the genome. *FAD2* also overlapped with a MP QTL (Sanyal and Linder, 2011). A plot of the 20 kb window around *FAD2* showed a pattern indicative of selection (Figure 3.8). *NADP-ME2* was associated with MP in both types of association tests and was also found within a MP QTL (Sanyal and Linder, 2011). *NADP-ME2* encodes a malic enzyme which decarboxylates malate into pyruvate, an early substrate in fatty acid biosynthesis (Wheeler *et al.*, 2005).

No-population-structure analyses

Eight lipid-related genes were identified by the no-population-structure test but

were not significant in the with-population-structure test. *BAP1 (AT3G61190)*, BON-associated protein, contained the most significant SNP in the genome and a BON protein, *BON2 (AT5G07300),* was also significant. *BAP1* and *BON2* both encode phospholipid binding proteins whose functions are not well understood. The function of *AT5G11140* is not known but is part of the phospholipase-like PEARLI 4 family of proteins and could therefore be involved in differential catabolism of FAs in seed oils. *GLX1 (AT1G11840)* encodes a glyxalase which is involved in the breakdown of lipids into sugars. *LACS9 (AT1G77590)* is a long chain acyl-CoA synthetase and was linked to three significant MP-associated SNPs. *CJD1 (AT1G08640)* was linked to four SNPs significantly associated with MP. A mutant of the *CJD1* chloroplast membrane protein has been implicated in changes to chloroplast fatty acid composition in leaves but has not yet been characterized in seeds (Ajjawi *et al.*, 2011). *ATPase E1-E2 (AT1G17500)* is responsible for phospholipid transport. *PDAT1 (AT5G13640)*, phospholipid:diacylglycerol acyltransferase 1, catalyzes the last acylation of the glycerol backbone to form the triacylglycerol (Zhang *et al.*, 2009).

Trans Regulatory Products

Many types of regulatory products were associated with MP variation. Of the 88 genes uniquely associated with MP, seven encode transcription factors, three are involved in DNA methylation, and two are involved in gene silencing. These regulatory elements are not known to play a role in lipid metabolism but a few were identified by multiple

tests as significantly associated with MP. The MP-associated SNP on chromosome 4, significant in both types of association tests and overlapped by a MP QTL, was located in a gene that encodes a transcription factor. *ESP4 (AT5G01400)* was significant in both association tests and encodes a protein involved in post-transcriptional gene silencing. One of the SNPs identified by the no-population-structure test and overlapped by QTL (Sanyal and Linder, 2011) but not by the with-population-structure test was linked to a gene that encodes transcription factor.

Genes Seemingly Unrelated to Lipids

The remaining 66 genes encode products with no known link to lipid metabolism or regulation (Table 3.3). Five of the 14 with-population-structure significant SNPs were linked to this category of genes. Two of these 5 were singleton SNPs that were not overlapped by QTL (Sanyal and Linder, 2011), so they might be false positives. Some false positives were expected in the top set of the no-population-structure genes. Nonetheless, some of these genes are linked to SNPs with such strong associations, they are likely to be true signals. For example, *DEGP7 (AT3G03380)* (a protease), contained 12 significant SNPs in the no-population-structure tests, while all other significant genes in this study were linked to four SNPs or less.

A priori candidate genes

When the full genome was analyzed, *FAD2* was the only *a priori* candidate gene significantly associated with MP in the genome-wide tests. When only the *a priori*

candidate gene SNPs were analyzed, *FATB (AT1G08510)*, was also significant, being linked with seven SNPs with an FDR<0.1. An additional five SNPs within *FAD2* were significant in this test.

Discussion

I found support for the hypothesis that seed oil melting point has been under selection to change with germination temperature (Linder, 2000). Specifically, spring temperatures are predictive of seed oil melting point in these spring-germinating accessions, particularly temperatures in February.

With respect to the genetic basis of the naturally occurring variation in seed oil MP, I identified several strong *a priori* and *a posteriori* candidate genes that have plausible influences on seed oil melting point. All of them either do or could change the ratios of the fatty acids with the greatest effects on melting point (16:0 and 20:1 to increase melting point, and 18:2 and 18:3 to decrease melting point). Further, many of the SNPs significantly associated with seed oil melting point in the GWA tests overlapped with melting point QTL (Sanyal and Randal Linder, 2011) (78.6% of the with-population-structure significant SNPs and 52.9% of the no-population-structure significant SNPs).

A priori candidate genes

Interestingly, the two *a priori* candidate genes linked to SNPs significantly associated with MP in this study, *FAD2* and *FATB*, have direct or indirect enzymatic roles in the synthesis of the four fatty acids with the strongest effects on seed oil melting point

(16:0, 18:2, 18:3, and 20:1). Variation in the activity level of variants of these enzymes could directly change the proportions of these fatty acids and the resulting seed oil melting point.

In the with-population-structure tests, *FAD2*, which catalyzes the desaturation of 18:1 to 18:2 (Okuley *et al.*, 1994), was linked to the most significant SNP in the genome. The rate of activity of *FAD2* could influence seed oil melting point directly by changing the proportion of 18:2, but also indirectly because 18:2 is the substrate for 18:3. 18:2 and 18:3 have the greatest effects on lowering seed oil melting point. The importance of *FAD2* for seed oil melting point is also supported by characterization of the seed oil composition of the *FAD2-1* (CS8041) ethyl methane sulfonate mutant (Lemieux *et al.*, 1990), which showed a 9.6 ºC increase its seed oil melting point (MP=22.5 ºC) relative to the wild-type (MP=12.9 ºC). The change in MP was caused by a 38% increase in 18:1, a 29% decrease in 18:2, a 15% decrease in 18:3, and a 10% decrease in 20:1.

FATB, an acyl-acyl carrier protein thioesterase which hydrolyzes acyl-ACP to produce free fatty acids has a substrate preference for 16:0-ACP (Salas and Ohlrogge, 2002) but will also hydrolyze 18:1-ACP (24% less activity) and 18:0-ACP (64% less activity). Because of this substrate preference for 16:0 and because 16:0 has the strongest influence on increasing seed oil melting point, down-regulation of *FATB* or protein variants of *FATB* could play a significant role in lowering *A. thaliana*'s melting point by decreasing the amount of 16:0 exported to the ER to be incorporated into seed oils. As

expected, a T-DNA insertion mutation in *FATB* (CS6525) (Sussman *et al.*, 2000) reduced the seed oil melting point by 3ºC as compared to wild-type. The MP reduction was produced by a 56% reduction of 16:0 in seeds and a 30% reduction of 18:0 (Bonaventure *et al.*, 2003).

Although only two of the fifteen *a priori* candidate genes were linked to significant SNPs, GWA mapping can fail to detect true associations due to complex genetic interactions such as epistasis (Chan *et al.*, 2010; Rowe *et al.*, 2008) and genetic heterogeneity (Bergelson and Roux, 2010). Many of the *a priori* candidate genes are part of gene families that might have redundant functions, making their individual influences difficult to identify through association mapping. There are five stearoyl-ACP desaturases (*ST1-5*), and they have high sequence similarity to *FAB2 (AT2G43710)* and *DES6 (AT1G43800)*. All seven genes likely contribute to production of 18:1 (Kachroo *et al.*, 2007). The four 3-ketoacyl-CoA synthases (*FAE1.1-1.4*), which perform elongation of the 18-carbon FAs to 20 and 22 carbons, are functionally related to 20 additional elongases with high sequence similarity (Blacklock and Jaworski, 2006). Finally, *FATA (AT3G25110)* has two copies (Kachroo *et al.*, 2007).

<u>*A posteriori* candidate genes</u>

With-population-structure test

Our understanding of the genetic drivers of complex traits is still in its infancy, so GWA studies should be expected often to provide new candidate genes rather than simply

confirm our *a priori* expectations. Of the 11 *a posteriori* candidate genes identified by the with-population-structure GWA study, none have been verified to affect seed oil melting point, but 4 of these genes (*ESP4*, *UAP56A*, *UAP56B*, and *NAC074*) encode products whose function could regulate lipid synthesis.

Polyadenylation has been shown to be involved in regulation of gene expression (Hunt *et al.*, 2008) and could be important in regulating lipid metabolism enzymes at the transcriptional level. *ESP4*, *UAP56A (AT5G11170)* and *UAP56B (AT5G11200)* may all be associated with the polyadenylation process. *ESP4* is a subunit of the cleavage and polyadenylation specificity factor which controls the action of the poly(A) polymerases for polyadenylation. A previous study that compared transgene expression in mutant *esp4-1* plants (and wild-type plants demonstrated a minimum of a 2-fold decrease in transgene mRNA and at least a 10-fold increase in transgene siRNA in the mutant (Herr *et al.*, 2006). *UAP56A* and *UAP56B* are DEAH RNA helicases, which are part of the polyadenylation machinery (Herr *et al.*, 2006). These results suggest that when melting point is under selective pressure, one path of evolution may proceed through moderating the rate of mRNA degradation. This would change the rate of enzyme production, altering fatty acid proportions and consequently leading to a change in the seed oil melting point.

Variation in trans-acting transcription factors that interact with the *a priori* candidate genes could contribute to melting point variation through differences in transcript abundance rather than variation in the enzymes themselves. *NAC074*

(AT4G28530), which is a transcription factor, may be acting in this fashion. Among the *a priori* and *a posteriori* candidate genes I identified, *NAC074*'s expression pattern is matched with only *FAD3*'s *(AT2G29980)* (Ruuska *et al.*, 2002). *FAD3*, catalyzes the desaturation of 18:2 to 18:3. In addition, *NAC074* is the only transcription factor identified in this study that has been shown previously to have a more than 2-fold expression increase during seed filling (Ruuska *et al.*, 2002). Finally, *NAC074* was significantly associated with seed oil melting point variation in both the with-population-structure and no-population-structure tests and was within a previously identified melting point QTL peak (Sanyal and Linder, 2011). It may be that *NAC074* has specificity to a small set of genes, including *FAD3*, and variation in its DNA binding domain may cause differences in its ability to promote initiation of transcription of *FAD3*.

No-population-structure test

Several of the top genes identified by the no-population-structure test encode products involved in lipid synthesis, binding or breakdown but not all of them have an obvious role in the modification of seed oil melting point. Two of these genes (*LACS9* and *PDAT1*) encode products with clear mechanisms by which they could modify seed oil composition, but they have not been functionally validated for their effects on that trait. They should therefore be candidate genes for additional molecular work to test their effect on seed oil melting point.

Lipid synthesis genes

LACS9, the major plastidial isoform responsible activating free fatty acids for export from the plastid, may indirectly act to lower seed oil melting point. It is highly expressed in developing seeds (Schnurr *et al.*, 2002) and has a substrate preference for 18:1 (Shockey *et al.*, 2002). Therefore, higher *LACS9* activity could lower seed oil melting point by increasing the substrate (18:1) available for *FAD2* thereby increasing the proportions of the low melting point fatty acids 18:2 and 18:3.

PDAT1 is an acyltransferase that catalyzes the addition of the final acyl group to the glycerol backbone during TAG production (Zhang *et al.*, 2009), and it has a substrate preference for polyunsaturated FAs (Bates *et al.*, 2009). This suggests that increased *PDAT1* activity could act to lower seed oil melting points by increasing the proportions of 18:2 and 18:3 incorporated into seed oils.

Conclusion

I identified a set of candidate genes that may have influenced the adaptive evolution of seed oil composition to optimize seed oil melting point relative to germination temperatures in nature. Knockouts of two of these genes, *FAD2* and *FATB*, have been shown to increase and decrease seed oil melting point, respectively. Characterization of the seed oil composition of up- and down-regulated transgenics and knockouts of the full set of candidate genes should be done to show whether they affect seed oil melting point in the predicted manner. To test whether these genes are adaptive,

fitness assays should be conducted comparing the relative fitness of the allelic variants in the same genetic background (either using NILs or transgenics) under cold and warm germination temperatures (Mackay, 2001; Wright and Gaut, 2005; Shindo *et al.*, 2007; Mitchell-Olds and Schmitt, 2006).

Chapter 4. Genetic basis of variation in fatty acid composition

Abstract

The renewable source of carbon provided by plant triacylglycerols fills an ever increasing demand for food, biodiesel and industrial chemicals. Each of these uses requires different compositions of fatty acid proportions in seed oils. Identifying the genes responsible for variation in seed oil composition in nature provides targets for bioengineering fatty acid proportions optimized for various industrial and nutrition goals. Here, I characterized the seed oil composition of 391 world-wide, wild accessions of *Arabidopsis thaliana* and performed a genome-wide association study of the 9 major fatty acids in the seed oil and 4 composite measures of the fatty acids. One hundred three SNPs were significantly associated with these traits, and 268 seed-expressed, *a posteriori* candidate genes were linked to them. Eight of these genes are involved in lipid metabolism. Four of the lipid metabolism genes confirm previous molecular work that demonstrated their effects on seed oil composition, while the remaining four would require validation. Thirty-four of the genes encode regulatory products, including transcription factors, kinases and methylases. The vast majority of the candidate genes have never been implicated in lipid metabolism and represent new targets for bioengineering.

Introduction

The triacylglycerols (TAGs) found in oilseeds provide a renewable source of highly reduced carbon for human consumption, industrial petrochemical replacement and biodiesel (Vanhercke et al., 2013, and refs therein). Currently, 160 million metric tonnes of plant oils are being produced globally, with 80% for food stuffs and 20% for industrial uses (Mielke (ed.), 2012) . With a growing population, dwindling petroleum supplies and the low carbon dioxide impact of biofuels, demand for plant oils will only increase. Many uses of seed oils require different optimal fatty acid compositions—the relative proportions of the fatty acids found in the seed oils. For example, for biodiesel, an ideal balance of high oxidative stability and improved cold tolerance results from high proportions of monounsaturated fatty acids, especially 18:1 (Durrett *et al.*, 2008). In canola, reduced levels of erucic acid and linolenic acid are desirable for human consumption (Yang *et al.*, 2012). On the other hand, oilseeds high in erucic acid (22:1) can be used to produce erucamide, a necessary component in the production of polyethylene and polypropylene films (Friedt and Luhs, 1998). To manipulate fatty acid composition efficiently for differing purposes, we must understand the interplay between the biosynthesis of fatty acids, their incorporation into TAGs, and the regulatory networks that control these processes.

While many of the enzymes involved in the synthesis and modification of fatty acids and their subsequent incorporation into TAGs have been well studied and are

conserved between most of the major oilseed species (e.g. *B. napus*, *A. thaliana*, castor bean and soybean) (Sharma and Chauhan, 2012), comparatively little is known about the structural and regulatory differences in these genes that produce such extensive variation in seed oil composition. The processes involved in fatty acid synthesis and incorporation into TAGs have been shown to be far more complex than a simple linear biochemical pathway (Chapman and Ohlrogge, 2012). Due to the identification of homologs of the lipid genes between many of the oilseed species, a more detailed understanding of how seed oil composition variation is produced within a single species in nature is likely to be useful for disentangling this process in other species.

Due to the availability of many tools for *Arabidopsis thaliana*, it has become the reference system for the study of lipid metabolism (Wallis and Browse, 2010). The value of studying lipid metabolism in *A. thaliana* is of additional interest because its fatty acid biosynthesis and regulation are similar to that found in *B.napus (Niu et al., 2009; Yang et al., 2012)*.

Recent work on seed oil synthesis and accumulation in *A. thaliana* has concentrated on elucidating the regulation of this process. The availability of seed-specific genome-wide expression data sets (M., Schmid *et al.*, 2005; Belmonte *et al.*, 2013) has led to several publications devoted to identifying coexpression networks involved in seed filling (Wang *et al.*, 2007; Peng and Weselake, 2011). To date, only a few transcription factors that specifically control seed storage accumulation have been

identified and validated (Wang *et al.*, 2007), (Baud *et al.*, 2007), (Mu *et al.*, 2008).

Three QTL mapping studies have attempted to uncover the genetic basis of variation in seed oil composition in *A. thaliana* (Hobbs *et al.*, 2004; O'Neill *et al.*, 2011; Sanyal and Randal Linder, 2011). Sixty-one genomic regions were identified, each being a large region covering dozens to hundreds of genes. In addition, while QTL mapping provides more power than many other mapping methods to detect true associations, it is limited to variation segregating between the parents. Among the QTL studies to date, only 10 pairs of accessions have been examined. Genes important for determining seed oil composition, therefore could easily have been missed.

These shortcomings of QTL mapping are largely overcome in genome-wide association studies (GWAS). Because GWAS uses large numbers of natural accessions, more of the genetic variation found in nature is represented and a much larger set of historical recombination events are reflected, allowing a larger number of genomic regions associated with the trait of interest to be identified. In addition, when performed with large numbers of markers, associations can be resolved to the genic level in most instances (Bergelson and Roux, 2010; Weigel, 2012). GWAS is now possible in species with a high degree of population structure, like *A. thaliana,* due to new statistical tests that control for relatedness (Kang *et al.*, 2008). On the other hand, GWAS suffers from a high number of false positives due to the large numbers of markers tested for association with trait variation. Recent work has shown that using expression data to filter GWAS

output can reduce false positive identification (Chan *et al.*, 2011).

My goal was to identify the genes responsible for variation in seed oil composition in *A. thaliana*. Using 391 accessions from across the geographic distribution of *A. thaliana*, I performed a GWAS on SNPs for the relative proportion of each of the nine major seed oil fatty acids (O'Neill *et al.*, 2003) and four biologically relevant composite traits. To reduce false positive associations, the genes linked to the significant SNPs were filtered for embryo-specific expression using publicly available expression data sets (Belmonte *et al.*, 2013). My results could be used to develop strategies for modifying seed oil composition for consumption or industrial purposes.

Materials and methods

Accessions and seed production

Three hundred ninety-one accessions (Figure 4.1) from across the geographic range of *A. thaliana* were chosen to capture as much of the phenotypic variation in seed oil composition as possible. Three hundred fifty-six were within the native range of *A. thaliana* (348 European, 6 Asian, 1 from the border between Europe and Asia, and 1 African), and 35 were recent introductions outside of its native range (United States 28, Japan 3, Cape Verde Islands 1, Canada 2, and New Zealand 1). Seeds were obtained from the Arabidopsis Biological Resource Center (ABRC) and the Juenger lab at the University of Texas at Austin.

To minimize environmental maternal effects, I produced seeds under common

garden conditions. Surface-sterilized seeds were soaked and stratified for six days at 4ºC prior to planting to break dormancy. Following stratification, plants were grown from November 2010 to March 2011 in a greenhouse with 16 hours of supplemental lighting from high pressure sodium, high-intensity discharge lights. Greenhouse temperatures were recorded every 4 minutes with an automatic data logger. Temperature ranged from a minimum of 6.1°C at night to a maximum of 28.9°C during the day with a mean of 18.1°C across all data points. Three replicates of each accession were planted in separate pots in a randomized block design for a total of 1,341 pots in 77 trays. An average of five seeds were planted in each pot and randomly thinned to one plant after germination and establishment. To decrease environmental variation, especially due to edge effects, pots were placed in every other tray slot such that every pot was equidistant. Twice per week the trays were subirrigated, fertilized with Dyna-Gro 7-9-5 plant food (Dyna-Gro Nutrition Solutions, Richmond, CA), and tray locations within the growing area were randomized. Once a plant bolted, it was isolated using the Arasystem (Betatech, Gent, BE). Seeds were harvested from each plant after the siliques matured and browned. Seeds were stored in coin envelopes at room temperature for 2 to 9 months before seed oil compositions were determined.

Seed oil composition analysis

Fatty acid compositions of the accessions were determined using gas chromatography of fatty acid methyl esters (FAMEs). For each of the three replicate plants per accession,

two extractions were performed and averaged on a per plant basis. Extraction procedures were modified from (Zheljazkov *et al.*, 2008): about thirty randomly chosen seeds per plant were crushed with a glass pestle in a 2mL glass autosampler vial and mixed with 150 μL of extraction buffer (75% hexane, 20% chloroform, and 5% 0.5M sodium methoxide in methanol). A minimum of five minutes elapsed before the first sample was injected into the chromatograph. Samples were analyzed using an HP 5890A gas chromatograph with a robotic autoloader and a DB-23 capillary column (Agilent Technologies, Santa Clara, CA). An initial oven temperature of 180°C was maintained for 5.5 min and then raised 15°C per min to a final temperature of 240°C for 0.5 min for a total run time of 10 min. Two injections were performed for every tenth sample to assess the repeatability of the results. In only 5 double-injections did the measured fatty acid proportions differ by more than 1%, with a maximum difference of 1.75%. Because of this high precision, only one of each of the double-injections was chosen at random to be used for the analyses. FAME peaks were visualized using a flame ionization detector and were identified by comparison to FAME standards 189-4, 189-5, 189-12 and 189-19 (Supelco, Bellefonte, PA). The standards were run before each set of seed oil extractions to calibrate retention times. Peaks were integrated using Agilent Chemstation software revision A.04.02.

A total of thirteen fatty acid traits were generated for each sample: the relative proportions of the nine principle fatty acids in *A. thaliana* seed oil (16:0, 18:0, 18:1, 18:2,

18:3, 20:0, 20:1, 20:2, and 22:1) (where the first number corresponds to the number of carbon atoms in the chain and the second, the number of double bonds), the sum of the proportions of the three fatty acids produced in the plastid (Plastid) (16:0, 18:0, and 18:1), the total proportion of saturated fatty acids (Sat) (16:0, 18:0, and 20:0), the total proportion of polyunsaturated fatty acids (PUFAs) (18:2, 18:3, and 20:2) and the total proportion of very long chain fatty acids (VLCFA) (20:0, 20:1, 20:2, and 22:1). The four composite traits were chosen for their potential importance in oil quality or genetic control of oil quality. Plastid fatty acids may be subject to maternal effects due to their location of synthesis. The other three traits (Sat, PUFA, VLCFA) are of interest for industrial and food applications (Vanhercke *et al.*, 2013).

The broad-sense heritability of each untransformed trait was determined using proc mixed in SAS. The class variable was accession and was considered a random variable. The calculations were also performed with tray as a random variable in the model and with box-cox transformed data.

Because the fatty acid biosynthetic pathway involves a complex network of reactions in which some of the fatty acids are precursors for the production of others, it is important to consider the correlations between their proportions in the seed oil. Pairwise Pearson correlations between the 9 fatty acids were performed with the cor.test function in R.

Association mapping

214,051 SNPs for *A. thaliana* were downloaded from http://walnut.usc.edu/2010/SNPs (Atwell *et al.*, 2010) (v3.06). Of these, I removed 13,748 minor alleles having frequencies of 5% or less (Kang *et al.*, 2008; Atwell *et al.*, 2010).

Three hundred ninety-one accessions were used to detect associations between SNP genotypes and variation in each of the 13 traits. Since using multiple measurements per accession can increase the power to detect associations (Kang *et al.*, 2008), all 3 biological replicates were used in the GWAS analyses. I used the EMMA.reml.t function of the R implementation EMMA v. 1.1.2 for Linux. EMMA identifies associations between SNPs and trait variation while controlling for genetic relatedness using a pairwise genetic similarity matrix (Kang *et al.*, 2008). The SNP genotype is modeled as a fixed effect and the genetic similarity matrix as a random effect. To meet the assumptions of EMMA, the traits were first subjected to box-cox transformations (Box and Cox, 1964) to approximate normal distributions.

Because association mapping using such a large dataset is computationally expensive and EMMA models each SNP independently, the dataset was divided into 1,020 sets of 210 SNPs each and run in parallel on the Lonestar cluster of the Texas Advanced Computing Center at the University of Texas at Austin. SNPs were considered significant at a -log(p-value)>5 and highly significant at -log(p-value)>10.

Identification of significant genes

Because linkage disequilibrium in *A. thaliana* decays within 10 kb on average (Kim *et al.*, 2007), all genes found up to 10 kb upstream or downstream of the significant SNP might be the causal variants. To filter this list of genes to a more plausible set affecting variation in seed oil composition, I used a publicly available dataset of global expression across 5 seed development stages in 7 subregions of *A. thaliana* accession Ws-0 (Belmonte *et al.*, 2013). The majority of triacylglycerol deposition in the seed occurs within the embryo proper from the torpedo to mature green cotyledon stages of development (Baud and Lepiniec, 2009). To capture genes that may be involved in initiation of this process, expression during the heart stage of development was also included. Only genes that had at least a moderate expression level of rma>=6 in the embryo proper during at least one of these stages (heart through mature cotyledon) were retained as putative *a posteriori* candidate genes. Gene functions and GO terms were obtained from TAIR (www.arabidopsis.org) and http://aralip.plantbiology.msu.edu/.

Previous analyses of seed-specific expression data sets has shown discrete patterns of expression of the genes involved in seed storage accumulation across the stages of seed development, suggesting control by dedicated sets of transcription factors (Wang *et al.*, 2007; Peng and Weselake, 2011). To search for transcription factors most likely to be regulating fatty acid synthesis and TAG accumulation, correlations between genes with transcription factor activity and those with lipid related functions were done

with the default settings of the cor function in R.

Results

Natural variation in fatty acid compositions

Among the 391 accessions, the nine major fatty acids found in *A. thaliana* seed oil displayed substantial variation in their relative proportions (Figure 4.2). Four of the unsaturated fatty acids (18:1, 18:2, 18:3, and 20:1) comprise an average of 83.4% of the seed oil, while the remaining 5 fatty acids are each less than 8%. 18:2 was the most abundant fatty acid in every accession (24.8% ± 1.7%, mean ± SD), and of the saturated fatty acids, 16:0 had the highest mean proportion (7.3% ± 0.55). In general, the fatty acids with higher mean proportions also had higher variances (Table 4.1). Coefficients of variation were calculated to see if the amount of variation in fatty acid proportions generally scaled with the means. In general, the least abundant fatty acids had the highest CV values and the most abundant had lower CV values indicating higher relative variation in the less abundant fatty acids. The highest CVs were found for 20:2 and 22:1 (CV = 14.2% and 12.0%, respectively) and the lowest for 18:2 (CV=6.9%) and 18:3 (CV=6.3%).

Heritabilities

The broad-sense heritability of each untransformed trait was high varying from 0.81 to 0.95 (Table 4.1), with the heritabilities of two of the commercially valuable composite traits, VLCFA and PUFA particularly high at 0.93. Models were also run with

tray as a random variable and using the box-cox transformed data. These modifications changed the heritabilities by no more than ±0.02 from the values reported.

Correlations between fatty acids

Thirty of the of the 36 fatty acid pairs were significantly correlated (p-value<0.05) and 28 of the 30 were highly correlated (p-value<0.01) (Table 4.2). Approximately half of the relationships between pairs of significantly correlated fatty acids were positive (13/30) and ranged from 0.127 to 0.562 with an average value of 0.334. The strongest positive correlation was between 20:1 and 22:1. This was a surprising result as a negative correlation was expected because 20:1 is the substrate for the production of 22:1. All saturated fatty acid pairs were positively correlated as were all but one pair (20:1/20:2) of the VLCFAs. The significant negative correlations were slightly stronger on average (-0.364, range: -0.113 to -0.766) than the significant positive correlations.

Of the 8 fatty acid pairs with direct substrate/product relationships (expected negative correlations), 2 pairs (18:1/20:1 and 18:2/18:3) were negatively correlated and five were positive (16:0/18:0, 18:0/18:1, 18:0/20:0, 18:2/20:2, and 20:1/22:1). One was non-significant (18:1/18:2). Interestingly, the strongest correlation (r=-0.766) was between the proportions of 18:1 and 20:2, which are not directly connected in the biochemical pathway, whereas 18:1 and its immediate product, 18:2, were not significantly correlated. In addition, the substrate-product pair, 18:2/20:2, was positively correlated (r=0.375).

Significant SNPs and the genes linked to them

After controlling for population structure, the mixed model EMMA identified a total of 103 SNPs significantly associated with at least 1 of the 13 traits (Figures 4.3 and 4.4). Of these, 54 were associated with only the relative proportions of the 9 fatty acids of the seed oils, 16 with only the composite traits, and 33 were shared between at least one the fatty acid proportion and one composite trait. Of the 87 fatty acid SNPs, 67 were associated with only one fatty acid, and the remaining 20 were associated with 2 or more fatty acids. The high level of fatty acid specificity of most of the SNPs was surprising given the highly correlated relationships among the fatty acid proportions. Also, even though the four composite traits (Sat, PUFA, VLCFA, and Plastid) were calculated using the individual fatty acid proportions, only 33 of 49 SNPs associated with them overlapped with those of at least one of the fatty acid proportions, suggesting that it may be an individual fatty acid of the composite trait driving the SNP associations. For the 16 SNPs associated with only composite traits, the collective effect of the fatty acids making up a composite trait caused SNP association.

Because linkage disequilibrium in *A. thaliana* decays within 10 kb on average (Kim *et al.*, 2007), all genes located between 10 kb upstream and downstream of each significant SNP were initially considered possible *a posteriori* candidate genes. The number of genes found within the 20 kb window ranged from 2 to 12, with an average of 7, for a total of 332 possible *a posteriori* candidate genes.

Expression information reduced the number of possible *a posteriori* candidate genes from 332 to 268. One hundred ninety genes met or exceeded the expression threshold and 78 were not included on the expression array and, therefore, could not be assessed so were retained as possible *a posteriori* candidates. All 103 SNPs were still linked to at least one *a posteriori* candidate gene after filtering for expression. Below I describe the genes of particular interest by category in descending order of significance.

Genes by category

The majority (58 of 103) of the significantly associated SNPs did not have genes with an obvious function in lipid metabolism or regulation of expression within 10 kb. Thirty-one SNPs had a total of 8 lipid metabolism genes within 10 kb, and 52 significant SNPs had a total of 34 trans-regulatory genes within 10 kb.

Lipid-related enzymes

Eight *a posteriori* candidates (*CJD1, AT1G09390, AT1G62610, KASIII, LPCAT2, KASII, FAD2,* and *FATA1*) encode products with functions known to be involved in lipid synthesis, degradation or transport. The most striking signal for these genes, and for the entire genome, is a cluster of 7 highly significant SNPs (-log(p-value) range from 10.1 to 33.2) found on chromosome 3. They were strongly associated with five traits (18:1, 18:2, 20:2, PUFA, and Plastid) (Figures 4.3 and 4.4). Of these 7 SNPs, 5 were located within *FAD2* (*AT3G12120*), and 1 was 1,435 bp downstream, while another was 31.9 kb downstream. *FAD2* encodes a fatty acid desaturase, which catalyzes the desaturation of

18:1 to 18:2.

Four of the remaining seven genes (*KASII, KASIII, AT1G62610* and *CJD1*) were associated with variation in the proportion of 16:0. *KASII* (*AT1G74960*) was also significantly associated with Sat. Three of the genes associations with 16:0 make sense. *KASII* encodes an enzyme which elongates 16:0-ACP to 18:0-ACP. *KASIII* (*AT1G62640*) and *AT1G62610*, which is a ketoacyl-ACP-reductase, are both part of the fatty acid synthase machinery required to produce 16:0 (Brown *et al.*, 2006). Not enough is known about *CJD1* to assess its association with 16:0. A mutant of the *CJD1* chloroplast membrane protein has been implicated in changes to chloroplast fatty acid composition in leaves but has not yet been characterized in seeds (Ajjawi *et al.*, 2011).

Differences in the relative proportions of 18:0 were associated with 2 lipid genes (*AT1G09390* and *FATA1*). *AT1G09390*, a lipase, was also associated with variation in total saturated fatty acid and encodes an enzyme with functional domains corresponding to lipase, acylhydrolase, and carboxylesterase activities. *FATA1* (*AT3G25110*) is a thioesterase, which hydrolyzes 18:1-ACP to allow export from the plastid.

Variation in the proportion of PUFA was associated with *LPCAT2* (*AT1G63050*) which produces an enzyme required for *PDAT1* to add PUFAs to TAGs (Xu *et al.*, 2012).

Trans Regulatory Genes

Thirty-four of the *a posteriori* candidate genes have trans-regulatory functions. Eleven encode transcription factors; 10 are involved in DNA methylation and/or gene

silencing; and 13 are kinases. None of these genes have been previously implicated in the regulation of lipid metabolism. Twenty-four regulatory genes were associated with only 1 trait.

In all of the cases of multiple trait associations, the traits have direct functional relationships. In 5 of the 10 multiple trait associations, more than 2 traits were associated. Transcription factor *AT3G12130*, 1,102 bp downstream of *FAD2*, is associated with the same traits as *FAD2* (18:1, 18:2, 20:2, plastid and PUFA). A kinase (*AT3G12200*) and a gene involved in RNA methylation (*AT3G11964*) were associated with 18:1, 20:2, plastid, and PUFA. Variation in 18:2, 20:2, and PUFA is associated with a transcription factor (*AT3G12730*) and a DNA methylation gene (*AT3G12710*).

Remaining Genes

Two hundred twenty-six *a posteriori* candidate genes are expressed in the embryo at appropriate stages but have no clear relationship to lipid synthesis, transport or regulation.

Correlations between transcription factors and lipid genes

The highly correlated expression patterns of the fatty acid synthesis genes suggests tight regulatory control at the transcriptional level. Of the 11 transcription factors associated with seed oil composition variation, 4 were not correlated with any of the 7 lipid genes on the expression array and are therefore less plausible as important components of those genes' regulation. Of the 7 remaining transcription factors, 4 were

significantly positively correlated with 3 lipid genes (*KASII*, *FAD2*, and *FATA1)* (r range: 0.79 to 0.96) and 1 was negatively correlated with these lipid genes (r range: -0.87 to -0.92).

Discussion

I determined the proportions of 9 individual fatty acids and 4 composite fatty acid traits in the seed oils of 391 natural accessions of *A. thaliana* and performed a genome-wide association study of each trait. Of the resulting 268 *a posteriori* candidate genes, 8 are lipid metabolism genes and 34 are regulatory genes, while the vast majority encode products with no clear functional role in lipid metabolism. The high heritabilities (H^2>0.8 in all cases) and extensive natural variation of the fatty acid proportions suggest selective breeding of many compositions would be possible in *A. thaliana*.

While *A. thaliana* is not a target for industrial or nutritional oil production, if the high heritabilities and variances found within this species are representative of closely related species, such as *B. napus*, then selective breeding may yet be a successful strategy. The least abundant fatty acids (18:0, 20:0, 20:2, and 22:1), while not capable of being bred to high proportions, can be manipulated to a greater extent than indicated by their low means, as shown by above average CV percentages. In addition to limitations on traditional breeding imposed by trait means and variances, breeding efforts may often be limited by the highly significant correlations between fatty acid pairs. On the other hand, in some cases, these correlations may facilitate breeding of desired compositions. For

example, the saturated fatty acids are positively correlated with one another, and it is often desirable to decrease total saturated fatty acids as a unit. The unexpected patterns of correlations I found between fatty acid pairs, such as positive or non-significant correlations between substrates and their products, support the idea of non-linearity of TAG production and the complexity of acyl fluxes (Vanhercke *et al.*, 2013).

Because the genes involved in fatty acid synthesis and TAG accumulation are conserved across many plant species, the genes associated with seed oil composition in *A. thaliana* provide likely candidates for QTL mapping and genetic engineering in non-model oilseed species (Sharma and Chauhan, 2012). For example, the functions of *FAD2* and *FAD3* were discovered and validated in *A. thaliana* and their DNA sequences were used to identify the specific gene copies of *FAD2* and *FAD3* in *B. napus* responsible for variation in the proportions of 18:1 and 18:3 *(Yang et al., 2012)*. This strategy could be applied to genes from my study to manipulate fatty acid proportions in other oilseed species. Recent advances in artificial microRNA technologies for seed-specific post-transcriptional gene silencing have been demonstrated to be highly effective in both model organisms and crop species (Sablok *et al.*, 2011; Belide *et al.*, 2012).

Lipid metabolism genes

The eight lipid metabolism genes associated with seed oil composition variation in this study represent good targets for industrial manipulation of fatty acid proportions. The associations identified in my study are in accordance with the effects on specific

fatty acid proportions of four lipid metabolism genes (*FAD2, KASII, FATA1,* and *LPCAT2*) engineered in other work. The remaining 4 lipid metabolism genes (*AT1G09390, KASIII, AT1G62610,* and *CJD1*) are known to contribute to fatty acid production but have not been shown to modify fatty acid proportions and, therefore, represent new putative targets for engineering specific oil compositions. A description of these eight genes, their associations, previous work and suggestions for engineering are discussed below.

Genes previously engineered to modify seed oil composition

FAD2 was tightly linked to SNPs that were 9 to 26 orders of magnitude more significantly associated with 5 traits (18:1, 18:2, 20:2, plastid, and PUFA) than any other SNPs linked to lipid-metabolism-associated genes. It encodes the enzyme that desaturates 18:1 to 18:2, which are the substrates for multiple downstream products (18:1 for 20:1, and 18:2 for 18:3 and 20:2). Molecular studies confirm the associations I found for *FAD2* in *A. thaliana* (Belide *et al.*, 2012) and show the same function for *FAD2* in soybean (Buhr *et al.*, 2002). With such a pivotal role in the fatty acid biosynthetic pathway, *FAD2* is an excellent target for manipulation of oil composition. It has, in fact, been used to bioengineer plants with modified compositions (Buhr *et al.*, 2002; Belide *et al.*, 2012). In *A. thaliana* and soybean, silencing of *FAD2* causes an increase in oleic acid proportions and a decrease in PUFAs.

KASII, which catalyzes the elongation of 16:0 to 18:0, was significantly

associated with variation in the proportions of total saturated fatty acids and 16:0. Earlier molecular work showed an increase in the proportion of 16:0 in seeds of *A. thaliana* through RNAi silencing of KASII (Pidkowich et al., 2007) and in a soybean *KASII* mutant (Head et al., 2012). Since oil high in 18:1 and low in saturated fatty acids is a desirable composition for biodiesel production (Durrett et al., 2008), I suggest that engineering of this composition may be possible through up-regulation of *KASII* in combination with up-regulation of *SAD,* which desaturates 18:0 to 18:1.

FATA1 encodes a thioesterase responsible for export of fatty acids from the plastid and was associated with variation in 18:0. The substrate preferences of 2 of the 3 thioesterases, *FATA1* and *FATB*, have been characterized (Durrett *et al.*, 2008). Both *FATA1* and *FATB* can hydrolyze 18:0 acyl-ACPs for export from the plastid but none of them displayed their highest activity with 18:0. Nonetheless, engineering low levels of 18:0 in TAGs by silencing *FATA1* may not be successful because reduced *FATA* activity has drastic effects on seed storage accumulation. Reduced thioesterase activity in a double *FATA* T-DNA mutant caused significant alterations in the proportions of most of the fatty acids in the seeds as well as reduced seed oil accumulation, which the authors suggest was due to a build up of acyl-ACPs in the plastid causing a negative feedback loop that shut down fatty acid production (Yang *et al.*, 2012).

Finally, the acyl transferase *LPCAT2* was associated with variation in PUFAs in this study, which is consistent with other work showing *LPCAT*s to have high activity in

plant species that accumulate large amounts of PUFAs in their TAGs (Friedt and Luhs, 1998). *LPCAT2* was recently found to participate in acyl editing by incorporating nascent fatty acids (mostly 18:1) into the membrane lipid phosphatidylcholine (Sharma and Chauhan, 2012). These fatty acids can then be further desaturated and subsequently transferred by *PDAT1* to the sn-3 position of a diacylglycerol to generate a TAG (Chapman and Ohlrogge, 2012). Therefore, by silencing both *LPCAT2* and *PDAT1*, it may be feasible to substantially decrease the proportion of PUFAs incorporated into TAGs.

Lipid genes newly implicated in seed oil composition variation

AT1G09390 (a lipase) was associated with variation in 18:0 and total saturated fatty acid proportions and may act by preferentially hydrolyzing TAGs with saturated fatty acids. Previous work has shown a decrease in lipid content in seeds during the final stages of maturation (Wallis and Browse, 2010), which is the stage of highest expression for this enzyme (Niu *et al.*, 2009; Yang *et al.*, 2012). If this mechanism of preferential degradation of TAGs with saturated fatty acid can be demonstrated in natural accessions, I suggest it may be possible to engineer a decrease in saturated fatty acid proportions in TAGs through up-regulation of this gene.

Both *KASIII* and *AT1G62610* (a ketoacyl-ACP reductase) were linked to SNPs associated with variation in 16:0, and both are critical parts of the fatty acid synthesis (FAS) complex (M., Schmid *et al.*, 2005; Belmonte *et al.*, 2013). While manipulation of

these enzymes may alter total seed oil content, it is unclear from current knowledge of their function and position in the pathway how they change oil composition.

CJD1 is a chloroplast membrane protein associated with proportions of 16:0 in seed TAGs. A T-DNA mutant of this gene has been shown to have altered leaf glycerolipid fatty acid composition in *A. thaliana* but its function in seeds is not known (Wang *et al.*, 2007; Peng and Weselake, 2011) because the seed oil composition of the mutant has not yet been characterized. It is unclear whether manipulation of this gene might affect seed oil fatty acid proportions.

Regulatory genes

None of the transcription factors previously implicated in the regulation of fatty acid production and seed oil accumulation (*FUS3* (Wang *et al.*, 2007), *LEC2* and *WRI1* (Baud *et al.*, 2007), and *LEC1* and *ABI3* (Mu *et al.*, 2008)) were associated with fatty acid proportions in this study, indicating that they may not be important for producing variation in seed oil composition in nature. While none of the 34 regulatory genes linked to SNPs associated with fatty acid proportions in this study have been shown to regulate seed oil composition, 2 transcription factors (*AT1G08620* and *AT3G56850*) and 3 kinases (*AT1G08650*, *AT2G35050* and *AT3G12200*) have been identified as part of either the fatty acid or oleosin coexpression network (Baud *et al.*, 2007). Using the embryo-specific global expression dataset of (Mu *et al.*, 2008), 7 of the 11 transcription factors associated with fatty acid proportions were correlated with expression of the lipid metabolism genes

I identified. To test the functional relationships of these 7 regulatory genes to seed oil composition, they should be silenced and their effect on seed oil composition characterized.

Remaining genes

While the remaining 226 genes identified by this study do not encode products with a clear functional relationship to lipid metabolism, they are significantly associated with variation in fatty acid proportions and are expressed in the embryo during the seed filling stages. Therefore, the set of genes affecting oil composition in *A. thaliana*, and possibly many other oilseed species, may be much broader than previously recognized, and these additional genes, if shown to affect oil composition could represent new targets for engineering seed oil composition and producing novel oils for consumption and industry.

The genes associated with natural variation in my study both confirm what is currently known about how seed oil composition is regulated and provide new candidates for bioengineering. My work indicates some food uses of oils might be best produced through traditional breeding. Because most industrial applications of plant lipids for petrochemical replacement will require high purity of the fatty acid of interest (sometimes in excess of 90%) and post-plant purification steps are costly (Vanhercke *et al.*, 2013), industrial applications would require genetic engineering to reach the desired purity.

Tables

Table 2.1 Effect of focal melting point type and density on days to emergence under cold and warm germination temperatures.

Effect	df[a]	F value	P value
Temperature=Cold			
Focal type	1, 307	2.04	0.154
Density	1, 307	1.52	0.219
Focal type*Density	1, 307	0.28	0.596
Temperature=Warm			
Focal type	1, 312	0.05	0.818
Density	1, 312	0.20	0.655
Focal type*Density	1, 312	0.00	0.959

[a]df stands for degrees of freedom: numerator, denominator

Table 2.2 Effect of Focal/Competitor type and density on fruit count under cold germination temperatures

Effect[a,b,c]	df	F value	P value
Focal/Competitor type (FC)	5, 299	51.82	**<0.0001**
Density	1, 299	139.38	**<0.0001**
Interaction: FC*Density	5, 299	0.65	0.6614
Contrast: H competition v none	1, 299	69.43	**<0.0001**
Contrast: L competition v none	1, 299	163.68	**<0.0001**
Contrast: FC LH v LL	1, 299	3.44	0.0645
Contrast: FC HL v HH	1, 299	6.09	**0.0142**

[a] H stands for high melting point lines
[b] L stands for low melting point lines
[c] FC stands for Focal/Competitor type, where the first letter is the focal type and the second letter is the competitor type

Table 2.3 Effect of Focal/Competitor type and density on fruit count under warm germination temperatures

Effect[a,b,c]	df	F value	P value
Focal/Competitor type	5, 304	44.35	**<0.0001**
Density	1, 304	67.31	**<0.0001**
Interaction: FC*Density	5, 304	0.32	0.8993
Contrast: H competition v none	1, 304	78.84	**<0.0001**
Contrast: L competition v none	1, 304	121.95	**<0.0001**
Contrast: FC LH v LL	1, 304	1.76	0.1853
Contrast: FC HL v HH	1, 304	6.28	**0.012**
Contrast: Density 1cm v 3cm	1, 304	67.31	**<0.0001**

[a] H stands for high melting point lines
[b] L stands for low melting point lines
[c] FC stands for Focal/Competitor type, where the first letter is the focal type and the second letter is the competitor type

Table 3.1 Climate and geographic regressions

Y[a]	X[b]	P-value[c]	r^2
MP	Latitude	0.101	0.00800
MP	Longitude	2.90e-07***	0.0780
MP	Altitude	0.0760	0.0100
MP	**February min temperature**	**8.40e-15*** **	**0.169**
MP	February mean temperature	6.61e-14***	0.159
MP	February max temperature	1.77e-12***	0.142
MP	March min temperature	3.24e-12***	0.139
MP	March mean temperature	2.68e-11***	0.128
MP	March max temperature	7.13e-10***	0.110
MP	April min temperature	3.71e-05***	0.0510
MP	April mean temperature	0.00160**	0.0300
MP	April max temperature	0.0360*	0.0130
MP	May min temperature	0.0400*	0.0130
MP	May mean temperature	0.706	0.000400
MP	May max temperature	0.267	0.00400

a MP is an abbreviation for seed oil melting point.
b The bold line is the most significant predictor of seed oil melting point.
c *p<0.05, **p<0.01, ***p<0.001

Table 3.2 Genes associated with seed oil melting point from the with-population structure tests

Gene[a]	TAIR ID	Distance[b]	All[c]	Native[d]	No PS[e]	QTL[f]
FAD2	AT3G12120	0	X	X		X
AT3G12210	AT3G12210	0	X			X
AT4G07515	AT4G07515	3986	X	X		
NAC074	AT4G28530	1542	X	X	X	X
ESP4	AT5G01400	0	X	X	X	
UAP56A	AT5G11170	0	X	X	X	X
UAP56B	AT5G11200	0	X	X	X	X
ASP3	AT5G11520	3207		X		X
NADP-ME2	AT5G11670	0	X	X	X	X
AGP15	AT5G11740	558	X	X	X	X
HB30	AT5G15210	0		X	X	X
AT5G28491	AT5G28491	6006	X	X		

a Genes in bold are *a priori* candidate genes.
b Distance between the gene and the closest significant SNP is given in bp. A value of zero indicates the SNP was within the gene.
c Significant using the All accessions dataset, N=391.
d Significant using the Native accessions dataset, N=356.
e Significant in the no-population-structure tests.
f Colocated with melting point QTL in (Sanyal and Randal Linder, 2011).

Table 3.3 Genes associated with melting point from the no-population structure tests

TAIR_ID	Distance[a]	All-noPS[b]	Native-noPS[c]	wPS[d]	QTL[e]
AT1G08640	0	X	X		
AT1G08910	0	X	X		
AT1G11840	1125	X	X		
AT1G12800	0		X		
AT1G13240	1287	X	X		
AT1G13270	0	X	X		
AT1G13290	0	X			
AT1G13330	984	X			
AT1G15757	1085	X	X		
AT1G15780	0	X			
AT1G15940	0	X	X		
AT1G15950	0	X	X		
AT1G17500	0	X	X		
AT1G18130	0	X	X		
AT1G18260	0	X	X		
AT1G18390	0	X			
AT1G18670	0	X	X		
AT1G18690	0	X	X		
AT1G18710	5249	X	X		
AT1G54330	0	X			X
AT1G68830	463	X	X		
AT1G69040	687	X	X		
AT1G77590	0	X	X		
AT1G78265	0		X		
AT1G78320	0	X	X		
AT2G31290	0	X	X		
AT2G31340	0	X	X		
AT2G31350	659	X	X		
AT2G31400	0		X		
AT2G31650	0		X		
AT2G31660	0	X	X		
AT2G31680	0	X	X		
AT2G31740	0	X	X		
AT2G34850	1014	X	X		
AT2G43800	0	X	X		
AT2G43810	581	X	X		
AT2G43820	0	X	X		
AT3G03350	0	X	X		X
AT3G03360	0	X	X		X
AT3G03380	0	X	X		X
AT3G03400	0	X	X		X
AT3G03405	0	X	X		X
AT3G03420	0	X	X		X

Table 3.3 Continued.

TAIR_ID	Distance	All-noPS	Native-noPS	wPS	QTL
AT3G10530	0	X	X		X
AT3G23230	2713		X		X
AT3G26890	0	X	X		
AT3G61190	0	X	X		
AT4G19650	0		X		
AT4G28530	1542	X	X	X	X
AT4G28690	0	X	X		X
AT4G28703	0	X	X		X
AT4G28706	0	X	X		X
AT4G29250	0	X	X		X
AT4G29580	0	X	X		X
AT4G32240	0	X	X		X
AT4G33810	1453		X		X
AT4G37720	1706	X			X
AT5G01400	0	X	X	X	
AT5G07300	0		X		X
AT5G07330	0	X	X		X
AT5G11110	0	X	X		X
AT5G11130	0	X	X		X
AT5G11140	0	X	X		X
AT5G11150	0	X	X		X
AT5G11170	0	X	X	X	X
AT5G11190	1224	X			X
AT5G11200	0	X	X	X	X
AT5G11210	0	X	X		X
AT5G11410	0	X	X		X
AT5G11490	0	X	X		X
AT5G11630	0	X	X		X
AT5G11650	0	X	X		X
AT5G11670	0	X	X	X	X
AT5G11700	0		X		X
AT5G11740	558		X	X	X
AT5G13640	0		X		X
AT5G15210	0		X	X	X
AT5G15920	0	X	X		X
AT5G15930	0	X	X		X
AT5G64190	0	X	X		
AT5G65600	851		X		
AT5G66440	2416	X			
AT5G66710	0	X	X		

a Distance between the gene and the closest significant SNP is given in bp. A value of zero indicates the SNP was within the gene.
b Significant using the All accessions dataset, N=391.
c Significant using the Native accessions dataset, N=356.
d Significant in the with-population-structure tests.
e Colocated with melting point QTL in (Sanyal and Randal Linder, 2011).

Table 4.1 Summary statistics of variation and heritability for fatty acid percentages in 391 accessions of *A. thaliana*

Trait	Mean	H^{2a}	Variance	SD[b]	CV (%)[c]
16:0	7.31	0.880	0.302	0.550	7.52
18:0	3.22	0.871	0.0836	0.289	8.98
18:1	16.0	0.841	2.25	1.50	9.36
18:2	24.8	0.940	2.89	1.70	6.85
18:3	20.1	0.809	1.60	1.27	6.28
20:0	2.12	0.829	0.0479	0.219	10.3
20:1	22.4	0.946	1.75	1.32	5.91
20:2	1.96	0.886	0.0772	0.278	14.2
22:1	2.02	0.846	0.0586	0.242	12.0
Plastid	26.6	0.849	2.63	1.62	6.10
PUFA	46.9	0.931	3.59	1.89	4.04
Sat	12.7	0.892	0.658	0.811	6.41
VLCFA	28.5	0.931	2.77	1.66	5.84

a Broad-sense heritability
b Standard deviation
c Coefficient of variation expressed as percentages

Table 4.2 Pearson correlations (r) between fatty acid pairs using the accession averages

	X16_0	X18_0	X18_1	X18_2	X18_3	X20_0	X20_1	X20_2
X18_0	**0.341*** **							
X18_1	**-0.163** **	**0.173*** **						
X18_2	**-0.366*** **	**-0.276*** **	-0.023					
X18_3	-0.031	**-0.174*** **	**-0.439*** **	**-0.335*** **				
X20_0	**0.217*** **	**0.502*** **	**-0.382*** **	**-0.392*** **	-0.026			
X20_1	**0.219*** **	0.025	**-0.318*** **	**-0.693*** **	-0.047	**0.484*** **		
X20_2	**-0.193*** **	**-0.373*** **	**-0.766*** **	**0.346*** **	**0.193*** **	**0.127***	0.074	
X22_1	**-0.113***	**-0.358*** **	**-0.619*** **	**-0.222*** **	**0.187*** **	**0.468*** **	**0.562*** **	**0.529*** **

*p<0.05, **p<0.01, ***p<0.001, significant pairs in bold.

Figures

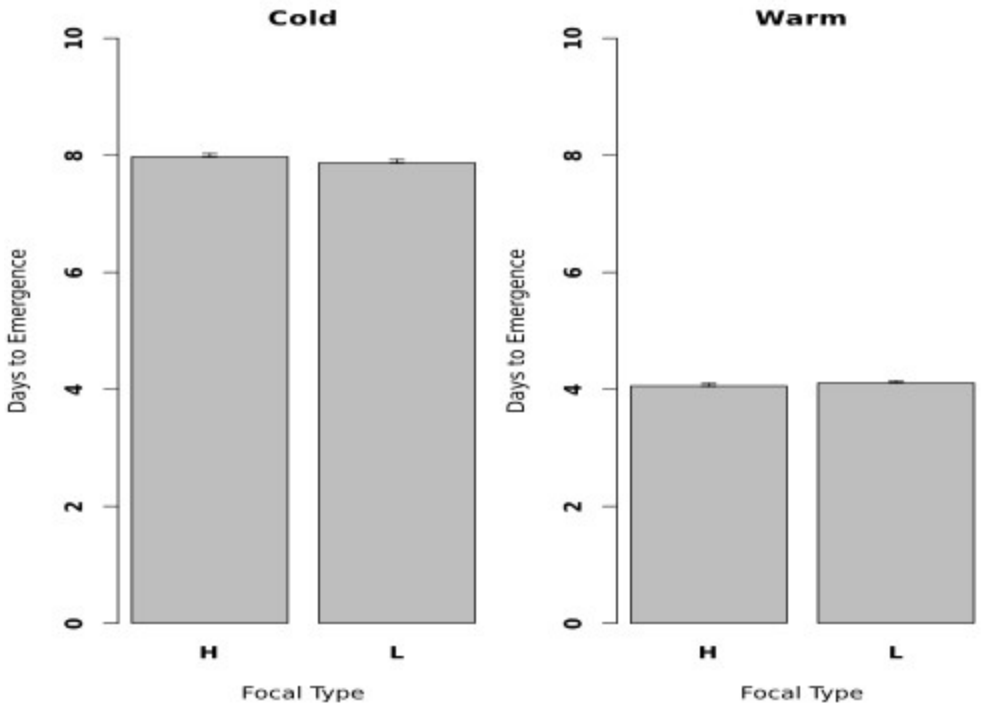

Figure 2.1 Mean and standard error of days to emergence for high (H) and low (L) melting point lines under cold and warm temperatures.

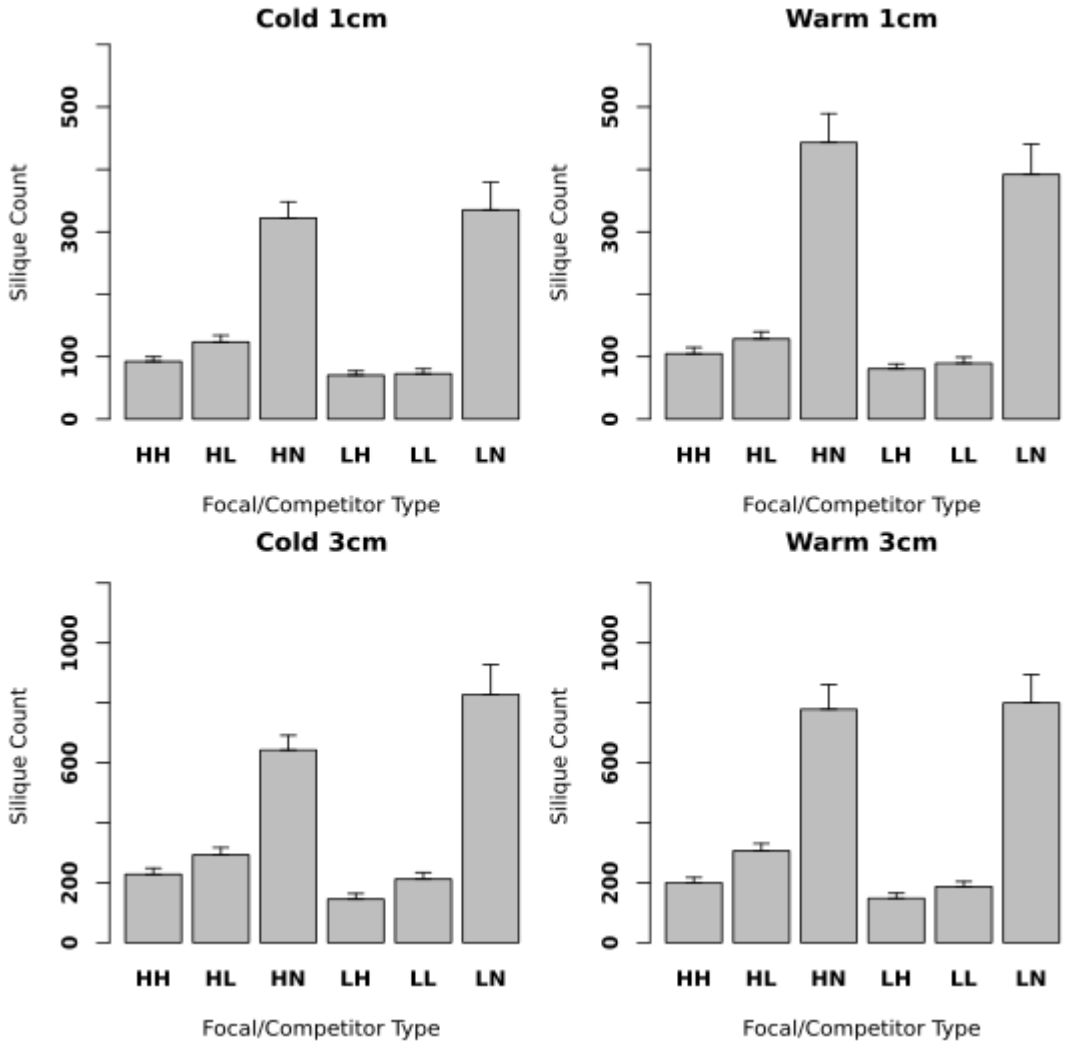

Figure 2.2 Mean and standard error of fruit count for each Focal/Competitor combination for each density (1cm or 3cm) under cold and warm temperatures.

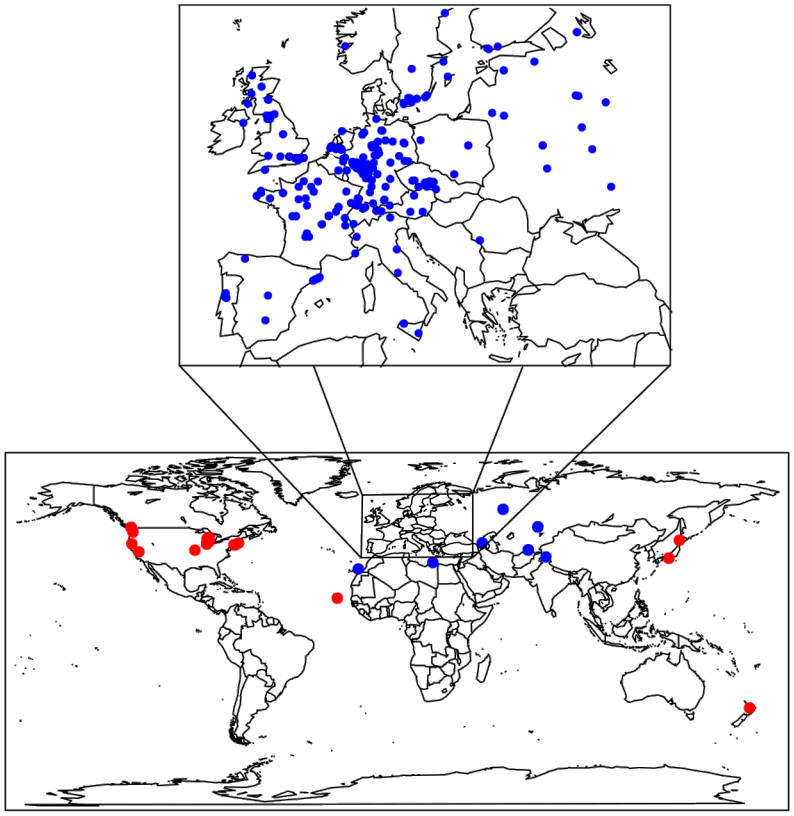

Figure 3.1 Geographic distribution of the 391 *A. thaliana* accessions used in this study. Blue dots represent native accessions (N=356) while red dots represent non-natives (N=35).

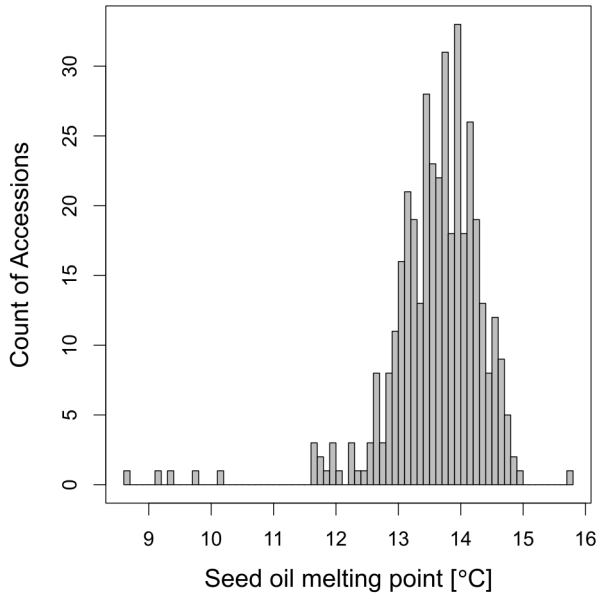

Figure 3.2 Distribution of seed oil melting points. Histogram of the average melting points for the accessions in the All dataset (N=391).

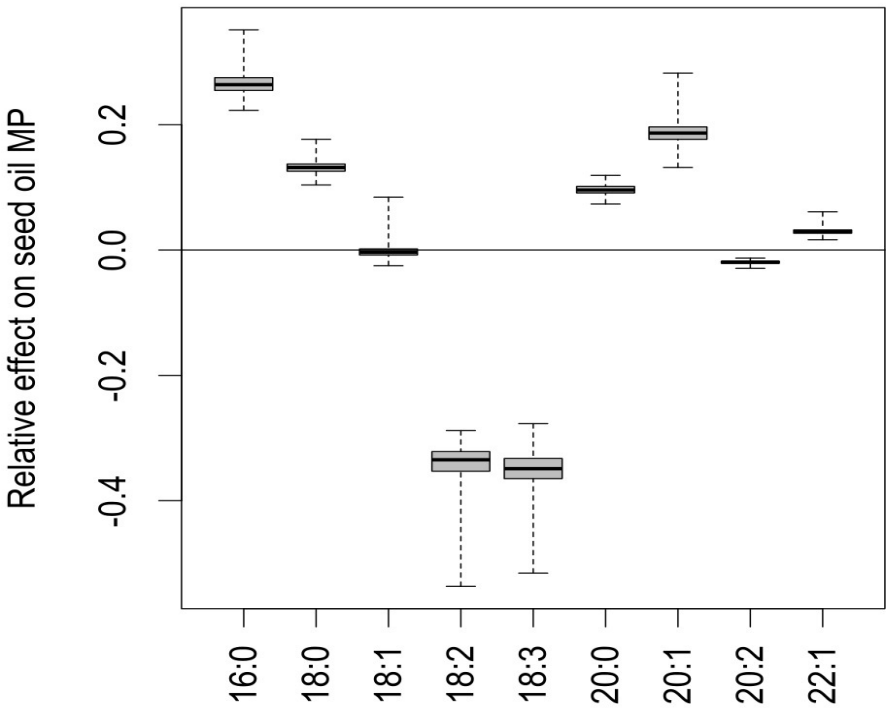

Figure 3.3 Relative effect of each fatty acid on seed oil melting point. Box plots of the effect each fatty acid had on seed oil melting point among the 391 accessions. A zero value means the fatty acid has no effect on seed oil melting point because its melting point is equivalent to the seed oil melting point.

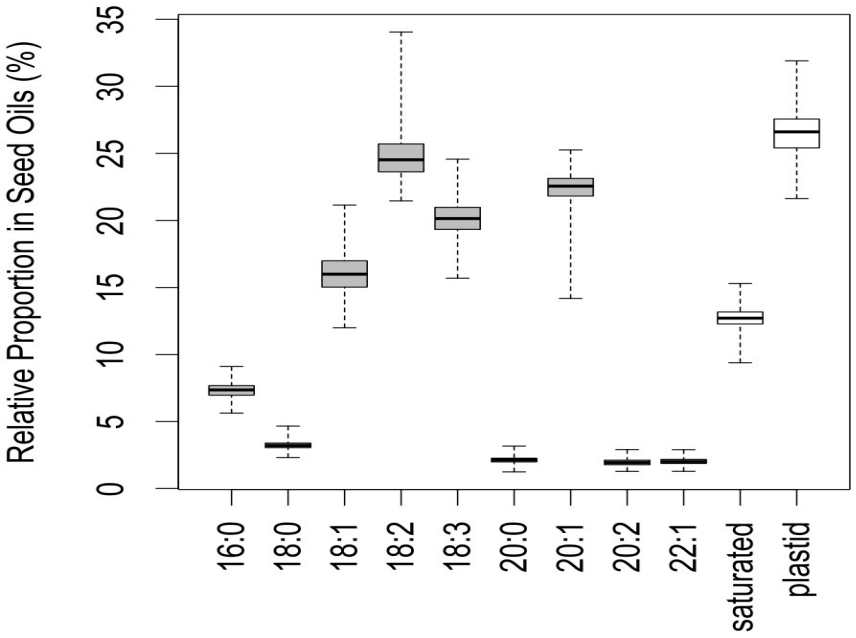

Figure 3.4 Relative proportions of the 9 major fatty acids in *A. thaliana* seeds and two integrated traits. Results are from the All accessions dataset (N=391).

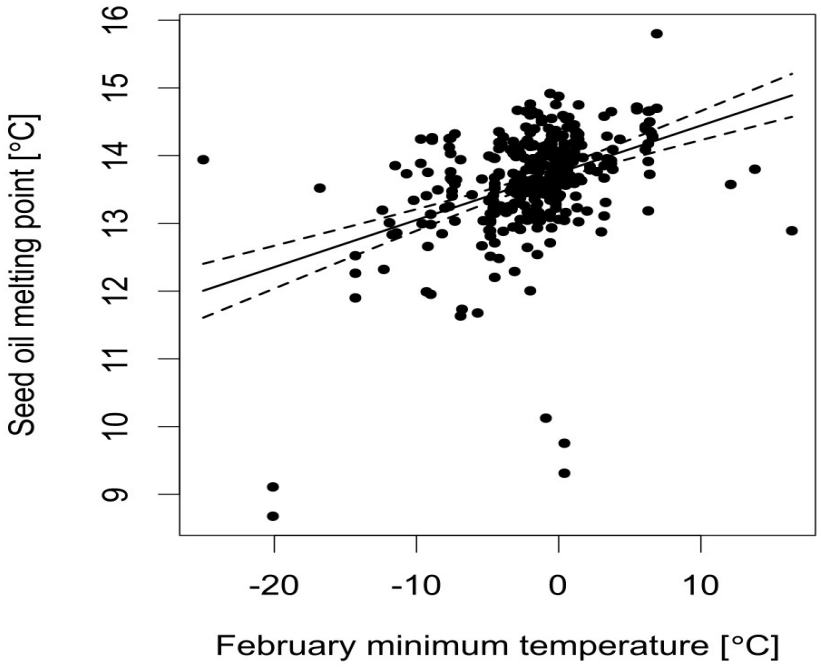

Figure 3.5 Melting point regressed on February minimum temperature. Best fit line and 95% confidence interval are shown ($r^2 = 0.169$).

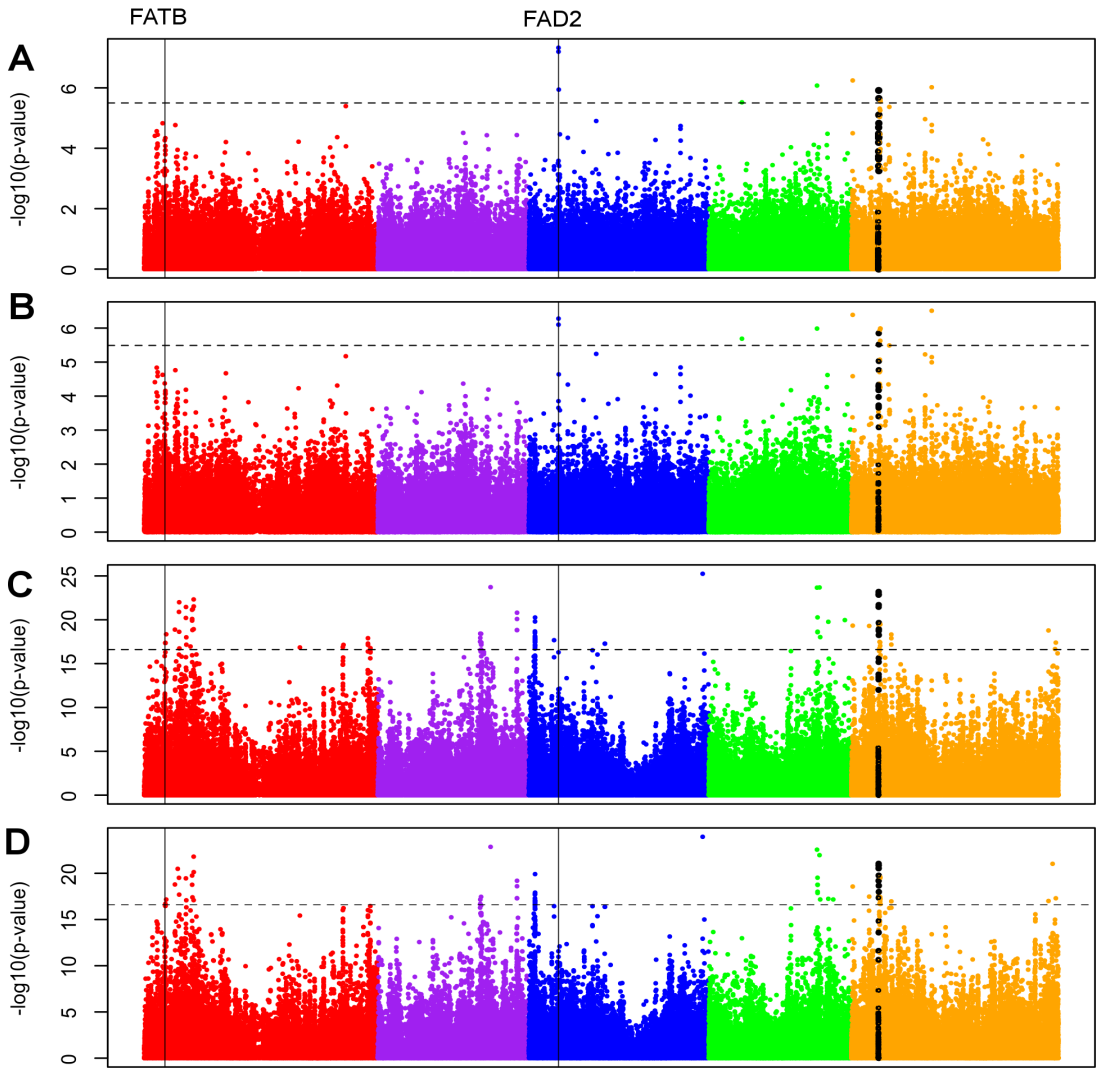

Figure 3.6 GWA results for seed oil melting point in *A. thaliana* using the EMMA package. (**A**) With-population-structure test using All accessions, N=391 (**B**) With-population-structure test using only Native accessions, N=356 (**C**) No-population-structure test using All accessions, N=391 (**D**) No-population-structure test using Native accessions, N=356. The the five chromosomes are represented by different colors. Horizontal dashed lines in the with-population-structure panels represent an FDR=0.05. Horizontal dashed lines in the no-population-structure panels show the cutoff boundaries for the top 100 SNPs. Vertical lines represent the *a priori* candidate genes significantly associated with seed oil melting point, *FATB* (AT1G08510) and *FAD2* (AT3G12120). The black points represent the 40kb region of extended association in chromosome 5.

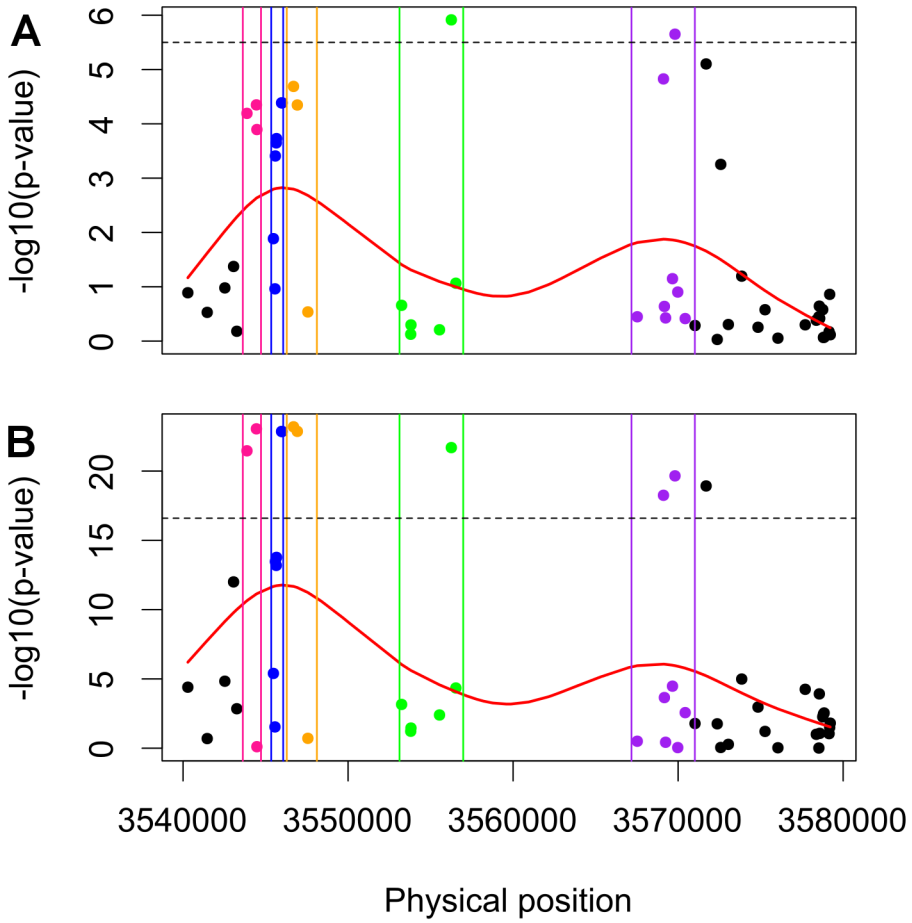

Figure 3.7 40kb region surrounding the extended region of association on Chromosome 5. (A) With-population-structure test using All accessions, N=391. Horizontal dashed line represents an FDR=0.05. (B) No-population-structure test using All accessions, N=391. Horizontal dashed line shows the boundary for the top 100 SNPs. Vertical lines represent the start and stop codons of genes with significant SNPs within their boundaries. SNPs and vertical lines are color coded by gene: AT5G11130 (pink), AT5G11140 (blue), AT5G11150 (orange), AT5G11170 (green) and AT5G11200 (purple). The red lines are smoothing splines showing the same underlying pattern of association in the two tests.

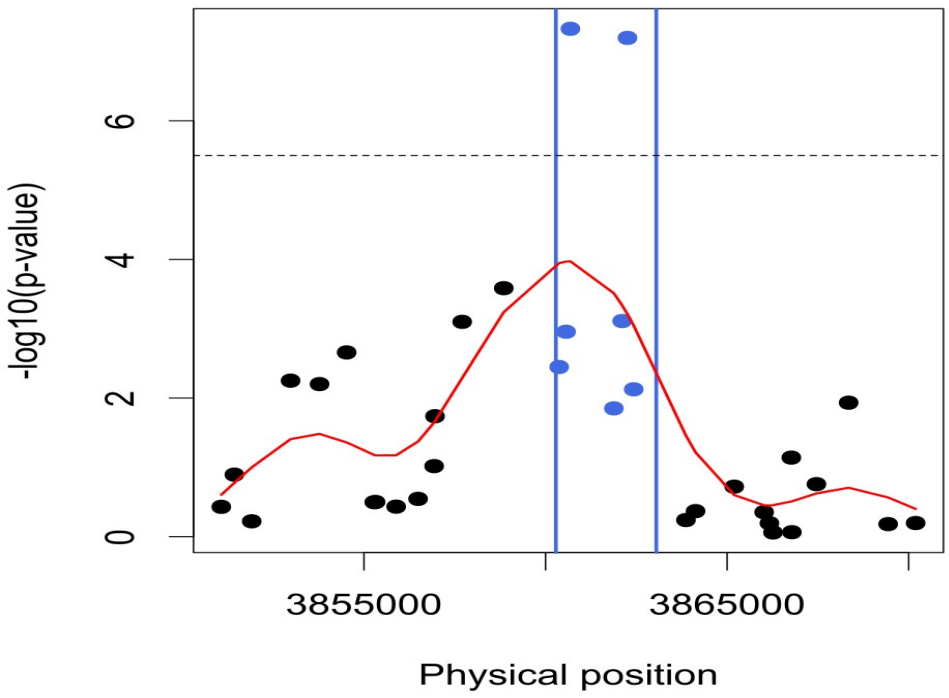

Figure 3.8 20 kb window surrounding the two most significant SNPs in the genome. Results are from the with-population-structure test using All accessions, N=391. Vertical lines represent the start and stop codons of the *a priori* candidate gene *FAD2* (AT3G12120). Points in blue are the SNPs in the coding region of *FAD2*. The red line depicts a smoothing spline showing an area of strong association that decreases with distance, a pattern indicative of selection.

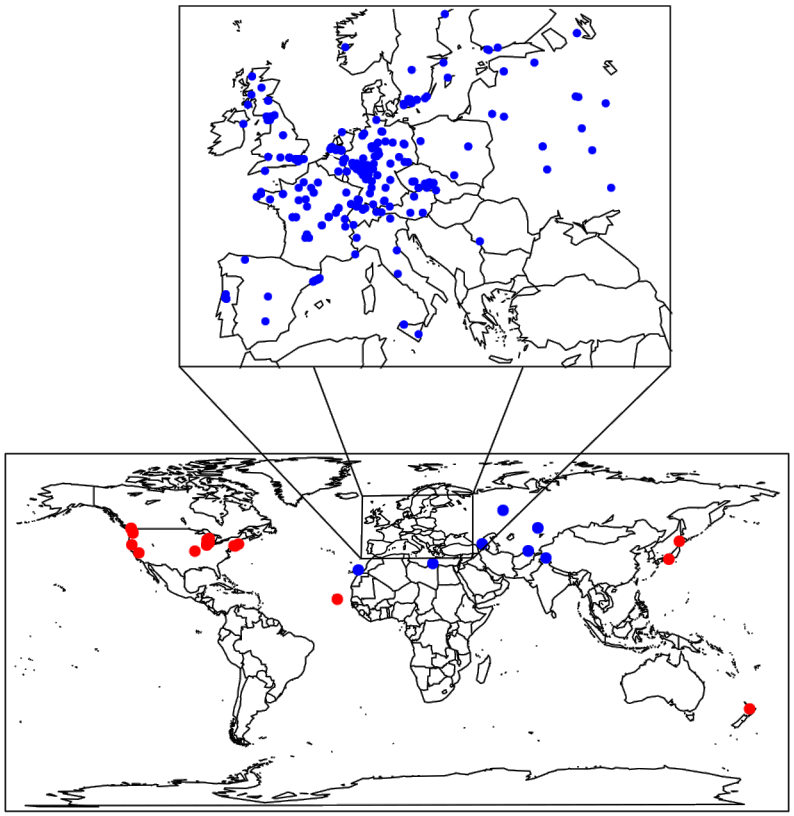

Figure 4.1 Geographic distribution of the 391 *A. thaliana* accessions used in this study. Blue dots represent native accessions (N=356) while red dots represent non-natives (N=35).

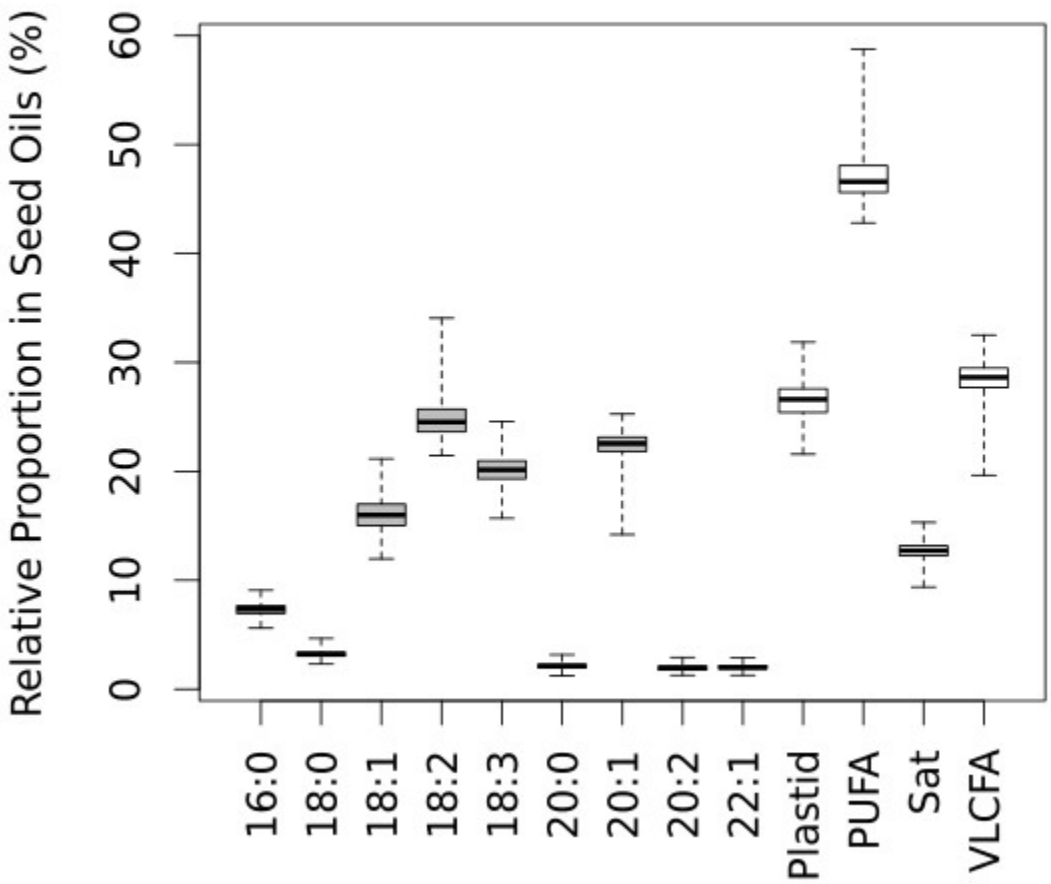

Figure 4.2 Relative proportions of the 9 major fatty acids in *A. thaliana* seeds using accession averages. (N=391)

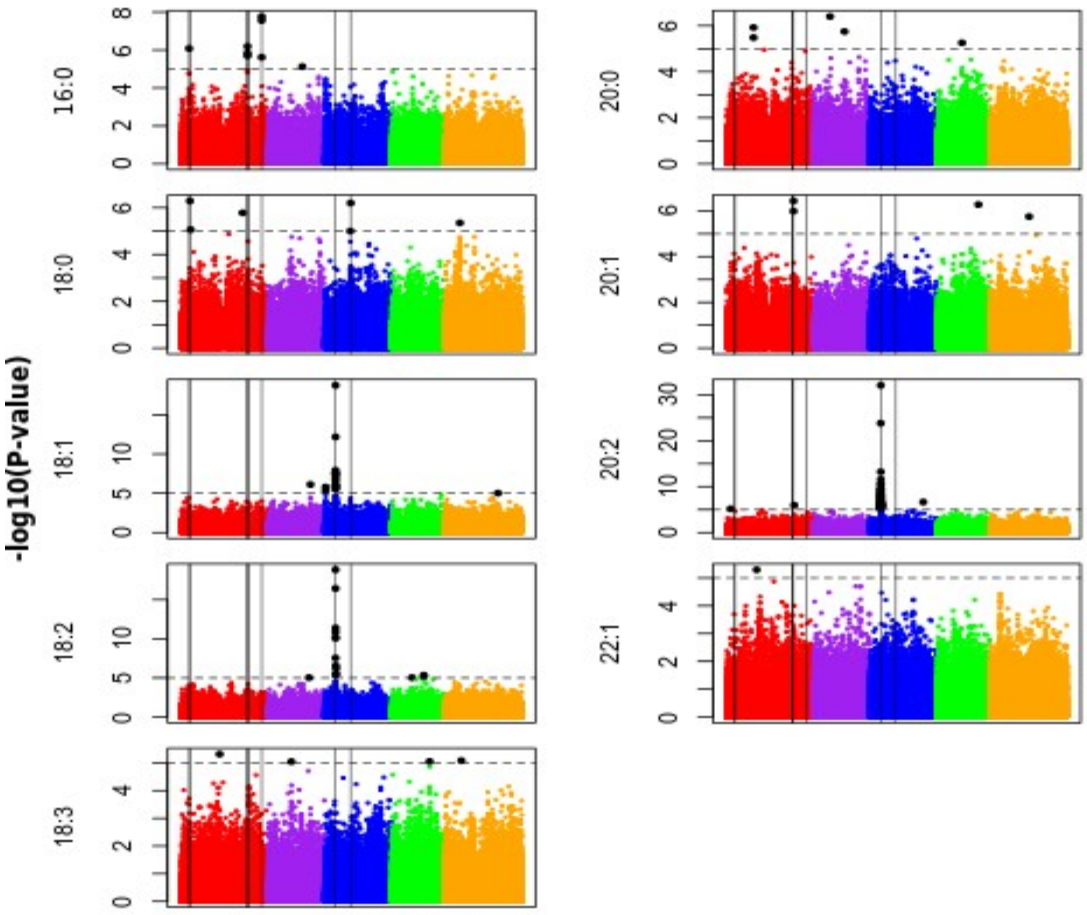

Figure 4.3 GWAS results for the 9 major fatty acids in A. thaliana using the EMMA package. The the five chromosomes are represented by different colors. Horizontal dashed lines show the significance threshold boundary of -log10(P-value)=5. Vertical lines represent the physical positions of the 8 lipid metabolism genes associated with fatty acid proportions. Some of the vertical lines are too close to be distinguished from one another and share a label (1=*CJD1/AT1G09390*, 2=*AT1G62610/KASIII/LPCAT2*, 3=*KASII*, 4=*FAD2*, 5=*FATA*).

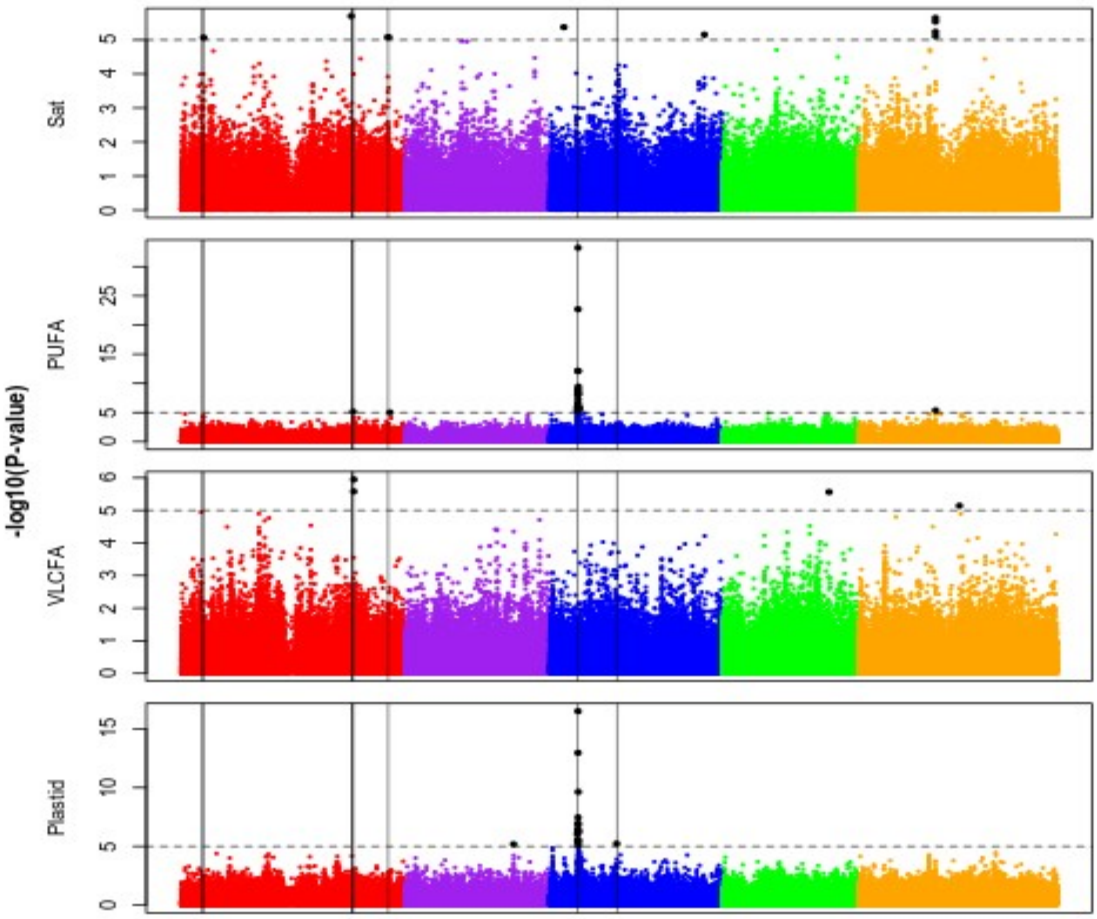

Figure 4.4 GWAS results for the 4 composite traits in A. thaliana using the EMMA package. The the five chromosomes are represented by different colors. Horizontal dashed lines show the significance threshold boundary of -log10(P-value)=5. Vertical lines represent the physical positions of the 8 lipid metabolism genes associated with fatty acid proportions. Some of the vertical lines are too close to be distinguished from one another and share a label (1=*CJD1/AT1G09390*, 2=*AT1G62610/KASIII/LPCAT2*, 3=*KASII*, 4=*FAD2*, 5=*FATA*).

References

Ajjawi, I., Coku, A., Froehlich, J.E., Yang, Y., Osteryoung, K.W., Benning, C. and Last, R.L. (2011) A J-like protein influences fatty acid composition of chloroplast lipids in Arabidopsis. *PLoS One*, **6**, e25368.

AL-MASKRI, A., Sajjad, M. and KHAN, S. (2012) Association Mapping: A Step Forward to Discovering New Alleles for Crop Improvement. *Int. J. ...*, **14**, 153–160.

Anastasio, A.E., Platt, A., Horton, M., Grotewold, E., Scholl, R., Borevitz, J.O., Nordborg, M. and Bergelson, J. (2011) Source verification of mis-identified Arabidopsis thaliana accessions. *Plant J.*, **67**, 554–66.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., *et al.* (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–31.

Bates, P.D., Durrett, T.P., Ohlrogge, J.B. and Pollard, M. (2009) Analysis of acyl fluxes through multiple pathways of triacylglycerol synthesis in developing soybean embryos. *Plant Physiol.*, **150**, 55–72.

Bates, P.D., Fatihi, A., Snapp, A.R., Carlsson, A.S., Browse, J. and Lu, C. (2012) Acyl editing and headgroup exchange are the major mechanisms that direct polyunsaturated fatty acid flux into triacylglycerols. *Plant Physiol.*, **160**, 1530–9.

Baud, S. and Lepiniec, L. (2009) Regulation of de novo fatty acid synthesis in maturing oilseeds of Arabidopsis. *Plant Physiol. Biochem.*, **47**, 448–55.

Baud, S., Mendoza, M.S., To, A., Harscoët, E., Lepiniec, L. and Dubreucq, B. (2007) WRINKLED1 specifies the regulatory action of LEAFY COTYLEDON2 towards fatty acid metabolism during seed maturation in Arabidopsis. *Plant J.*, **50**, 825–38.

Beck, J.B., Schmuths, H. and Schaal, B. a (2008) Native range genetic variation in Arabidopsis thaliana is strongly geographically structured and reflects Pleistocene glacial dynamics. *Mol. Ecol.*, **17**, 902–15.

Belide, S., Petrie, J.R., Shrestha, P. and Singh, S.P. (2012) Modification of Seed Oil Composition in Arabidopsis by Artificial microRNA-Mediated Gene Silencing. *Front. Plant Sci.*, **3**, 168.

Belmonte, M.F., Kirkbride, R.C., Stone, S.L., *et al.* (2013) Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, E435–44.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Bergelson, J. and Roux, F. (2010) Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nat. Rev. Genet.*, **11**, 867–79.

Blacklock, B.J. and Jaworski, J.G. (2006) Substrate specificity of Arabidopsis 3-ketoacyl-CoA synthases. *Biochem. Biophys. Res. Commun.*, **346**, 583–90.

Bonaventure, G., Salas, J., Pollard, M.R. and Ohlrogge, J.B. (2003) Disruption of the FATB gene in Arabidopsis demonstrates an essential role of saturated fatty acids in plant growth. *Plant Cell ...*, **15**, 1020–1033.

Box, G. and Cox, D. (1964) An analysis of transformations. *J. R. Stat. Soc. Ser. B*, **26**, 211–252.

Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., Bergelson, J., Cuguen, J. and Roux, F. (2010) Linkage and association mapping of Arabidopsis thaliana flowering time in nature. *PLoS Genet.*, **6**, e1000940.

Brown, A.P., Affleck, V., Fawcett, T. and Slabas, A.R. (2006) Tandem affinity purification tagging of fatty acid biosynthetic enzymes in Synechocystis sp. PCC6803 and Arabidopsis thaliana. *J. Exp. Bot.*, **57**, 1563–71.

Buhr, T., Sato, S., Ebrahim, F., Xing, A., Zhou, Y., Mathiesen, M., Schweiger, B., Kinney, A. and Staswick, P. (2002) Ribozyme termination of RNA transcripts down-regulate seed fatty acid genes in transgenic soybean. *Plant J.*, **30**, 155–63.

Cai, Z.Q., Jiao, D.Y., Tang, S.X., Dao, X.S., Lei, Y.B. and Cai, C.T. (2012) Leaf Photosynthesis, Growth, and Seed Chemicals of Sacha Inchi Plants Cultivated Along an Altitude Gradient. *Crop Sci.*, **52**, 1859.

Cao, J., Schneeberger, K., Ossowski, S., *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.*, **43**, 956–63.

Cernac, A. and Andre, C. (2006) WRI1 is required for seed germination and seedling establishment. *Plant Physiol.*, **141**, 745–757.

Chan, E.K.F., Rowe, H.C., Corwin, J. a, Joseph, B. and Kliebenstein, D.J. (2011) Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana. *PLoS Biol.*, **9**, e1001125.

Chan, E.K.F., Rowe, H.C. and Kliebenstein, D.J. (2010) Understanding the evolution of defense metabolites in Arabidopsis thaliana using genome-wide association mapping. *Genetics*, **185**, 991–1007.

Chapman, K.D. and Ohlrogge, J.B. (2012) Compartmentation of triacylglycerol accumulation in plants. *J. Biol. Chem.*, **287**, 2288–94.

Childs, L.H., Witucka-Wall, H., Günther, T., Sulpice, R., Korff, M. V, Stitt, M., Walther, D., Schmid, K.J. and Altmann, T. (2010) Single feature polymorphism (SFP)-based selective sweep identification and association mapping of growth-related metabolic traits in Arabidopsis thaliana. *BMC Genomics*, **11**, 188.

Debieu, M., Tang, C., Stich, B., *et al.* (2013) Co-variation between seed dormancy, growth rate and flowering time changes with latitude in Arabidopsis thaliana. *PLoS One*, **8**, e61075.

Donohue, K., Dorn, L., Griffith, C., Kim, E., Aguilera, A., Polisetty, C.R. and Schmitt, J. (2005) Niche construction through germination cueing: life-history responses to timing of germination in Arabidopsis thaliana. *Evolution*, **59**, 771–85.

Durrett, T.P., Benning, C. and Ohlrogge, J. (2008) Plant triacylglycerols as feedstocks for the production of biofuels. *Plant J.*, **54**, 593–607.

Dyer, a. R., Fenech, a. and Rice, K.J. (2000) Accelerated seedling emergence in interspecific competitive neighbourhoods. *Ecol. Lett.*, **3**, 523–529.

Eckey, E.W. (1954) *Vegetable fats and oils.*, Reinhold, New York.

Friedt, W. and and Luhs, W. (1998) Recent developments and perspectives of industrial rapeseed breeding. *Fett/Lipid*, **100**, 219–226.

Geber, M. and Griffen, L. (2003) Inheritance and natural selection on functional traits. *Int. J. Plant Sci.*, **164**.

Hall, D., Tegström, C. and Ingvarsson, P.K. (2010) Using association mapping to dissect the genetic basis of complex traits in plants. *Brief. Funct. Genomics*, **9**, 157–65.

Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C., Roux, F. and Bergelson, J. (2011) Adaptation to climate across the Arabidopsis thaliana genome. *Science*, **334**, 83–6.

Harwood, J.L. (1980) Plant acyl lipids: Structure, distribution, and analysis. In P. K. Stumpf, ed. *The biochemistry of plants: a comprehensive treatise.* Academic Press, New York, pp. 2–55.

Herr, A.J., Molnàr, A., Jones, A. and Baulcombe, D.C. (2006) Defective RNA processing enhances RNA silencing and influences flowering of Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 14994–5001.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. and Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.*, **25**, 1965–1978.

Hobbs, D., Flintham, J. and Hills, M. (2004) Genetic control of storage oil synthesis in seeds of Arabidopsis. *Plant Physiol.*, **136**, 3341–3349.

Huang, X., Schmitt, J., Dorn, L., Griffith, C., Effgen, S., Takao, S., Koornneef, M. and Donohue, K. (2010) The earliest stages of adaptation in an experimental plant population: strong selection on QTLS for seed dormancy. *Mol. Ecol.*, **19**, 1335–51.

Hunt, A.G., Xu, R., Addepalli, B., *et al.* (2008) Arabidopsis mRNA polyadenylation machinery: comprehensive analysis of protein-protein interactions and gene expression profiling. *BMC Genomics*, **9**, 220.

Kachroo, A., Shanklin, J., Whittle, E., Lapchyk, L., Hildebrand, D. and Kachroo, P. (2007) The Arabidopsis stearoyl-acyl carrier protein-desaturase family and the contribution of leaf isoforms to oleic acid synthesis. *Plant Mol. Biol.*, **63**, 257–71.

Kang, H.M., Zaitlen, N. a, Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J. and Eskin, E. (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–23.

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D. and Nordborg, M. (2007) Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nat. Genet.*, **39**, 1151–5.

Lasky, J.R., Marais, D.L. Des, McKay, J.K., Richards, J.H., Juenger, T.E. and Keitt, T.H. (2012) Characterizing genomic variation of Arabidopsis thaliana: the roles of geography and climate. *Mol. Ecol.*, **21**, 5512–29.

Lehninger, A.I. (1993) Biochemistry. In Worth, New York.

Lemieux, B., Miquel, M., Somerville, C.R. and Browse, J. (1990) Mutants of Arabidopsis with alterations in seed lipid fatty acid composition. *TAG Theor.*, **80**, 234–240.

Li, Y., Huang, Y., Bergelson, J., Nordborg, M. and Borevitz, J.O. (2010) Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 21199–204.

Li-Beisson, Y., Shorrosh, B., Beisson, F., *et al.* (2010) Acyl-lipid metabolism. *Arabidopsis Book*, **8**, e0133.

Li-Beisson, Y., Shorrosh, B., Beisson, F., *et al.* (2013) Acyl-lipid metabolism. *Arabidopsis Book*, **11**, e0161.

Linder, C. (2000) Adaptive evolution of seed oils in plants: accounting for the biogeographic distribution of saturated and unsaturated fatty acids in seed oils. *Am. Nat.*, **156**, 442–458.

Lister, C. and Dean, C. (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in Arabidopsis thaliana. *Plant J.*

Mackay, T. (2001) Quantitative trait loci in Drosophila. *Nat. Rev. Genet.*, **2**, 1–10.

Malkin, T. (1954) The polymorphism of glycerides. In and T. M. R.T. Holman, W.O. Lundberg, ed. *Progress in the chemistry of fats and other lipids*. Pergamon, New York, pp. 1–50.

Mauricio, R. (2001) Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nat. Rev. Genet.*, **2**, 370–81.

Mercer, K.L., Alexander, H.M. and Snow, A. a (2011) Selection on seedling emergence timing and size in an annual plant, Helianthus annuus (common sunflower, Asteraceae). *Am. J. Bot.*, **98**, 975–85.

Millar, a a and Kunst, L. (1999) The natural genetic variation of the fatty-acyl composition of seed oils in different ecotypes of Arabidopsis thaliana. *Phytochemistry*, **52**, 1029–33.

Miller, T., Winn, A. and Schemske, D. (1994) The effects of density and spatial distribution on selection for emergence time in Prunella vulgaris (Lamiaceae). *Am. J. Bot.*, **81**, 1–6.

Miquel, M.F. and Browse, J. a. (1994) High-Oleate Oilseeds Fail to Develop at Low Temperature. *Plant Physiol.*, **106**, 421–427.

Mitchell-Olds, T. and Schmitt, J. (2006) Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature*, **441**, 947–52.

Mu, J., Tan, H., Zheng, Q., *et al.* (2008) LEAFY COTYLEDON1 is a key regulator of fatty acid biosynthesis in Arabidopsis. *Plant Physiol.*, **148**, 1042–54.

Myles, S., Peiffer, J., Brown, P.J., Ersoz, E.S., Zhang, Z., Costich, D.E. and Buckler, E.S. (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell*, **21**, 2194–202.

Nemri, A., Atwell, S., Tarone, A.M., Huang, Y.S., Zhao, K., Studholme, D.J., Nordborg, M. and Jones, J.D.G. (2010) Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 10302–7.

Niu, Y., Wu, G.-Z., Ye, R., *et al.* (2009) Global analysis of gene expression profiles in Brassica napus developing seeds reveals a conserved lipid metabolism regulation with Arabidopsis thaliana. *Mol. Plant*, **2**, 1107–22.

Nordborg, M., Hu, T.T., Ishino, Y., *et al.* (2005) The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol.*, **3**, e196.

O'Neill, C.M., Gill, S., Hobbs, D., Morgan, C. and Bancroft, I. (2003) Natural variation for seed oil composition in Arabidopsis thaliana. *Phytochemistry*, **64**, 1077–1090.

O'Neill, C.M., Morgan, C., Hattori, C., *et al.* (2011) Towards the genetic architecture of seed lipid biosynthesis and accumulation in Arabidopsis thaliana. *Heredity (Edinb).*, **108**, 115–23.

Okuley, J., Lightner, J., Feldmann, K., Yadav, N., Lark, E. and Browse, J. (1994) Arabidopsis FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. *Plant Cell ...*, **6**, 147–158.

Orr, H.A. (2005) The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.*, **6**, 119–27.

Ostrowski, M.-F., David, J., Santoni, S., *et al.* (2006) Evidence for a large-scale population structure among accessions of Arabidopsis thaliana: possible causes and consequences for the distribution of linkage disequilibrium. *Mol. Ecol.*, **15**, 1507–17.

Peng, F.Y. and Weselake, R.J. (2011) Gene coexpression clusters and putative regulatory elements underlying seed storage reserve accumulation in Arabidopsis. *BMC Genomics*, **12**, 286.

Picó, F.X., Méndez-Vigo, B., Martínez-Zapater, J.M. and Alonso-Blanco, C. (2008) Natural genetic variation of Arabidopsis thaliana is geographically structured in the Iberian peninsula. *Genetics*, **180**, 1009–21.

Pidkowich, M.S., Nguyen, H.T., Heilmann, I., Ischebeck, T. and Shanklin, J. (2007) Modulating seed beta-ketoacyl-acyl carrier protein synthase II level converts the composition of a temperate seed oil to that of a palm-like tropical oil. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 4742–7.

Platt, A., Horton, M., Huang, Y.S., *et al.* (2010) The scale of population structure in Arabidopsis thaliana. *PLoS Genet.*, **6**, e1000843.

Rafalski, J.A. (2010) Association genetics in crop improvement. *Curr. Opin. Plant Biol.*, **13**, 174–80.

Rees, M. and Long, M. (1992) Germination biology and the ecology of annual plants. *Am. Nat.*, **139**, 484–508.

Rowe, H.C., Hansen, B.G., Halkier, B.A. and Kliebenstein, D.J. (2008) Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome. *Plant Cell*, **20**, 1199–216.

Ruuska, S., Girke, T., Benning, C. and Ohlrogge, J.B. (2002) Contrapuntal networks of gene expression during Arabidopsis seed filling. *Plant Cell*, **14**, 1191–1206.

Sablok, G., Pérez-Quintero, A.L., Hassan, M., Tatarinova, T. V and López, C. (2011) Artificial microRNAs (amiRNAs) engineering - On how microRNA-based silencing methods have affected current plant silencing research. *Biochem. Biophys. Res. Commun.*, **406**, 315–9.

Salas, J. and Ohlrogge, J. (2002) Characterization of substrate specificity of plant FatA and FatB acyl-ACP thioesterases. *Arch. Biochem. Biophys.*, **403**, 25–34.

Sanyal, A. and Linder, C. (2013) Plasticity and constraints on fatty acid composition in the phospholipids and triacylglycerols of Arabidopsis accessions grown at different temperatures. *BMC Plant Biol.*, **13**, 0–13.

Sanyal, A. and Randal Linder, C. (2011) Quantitative trait loci involved in regulating seed oil composition in Arabidopsis thaliana and their evolutionary implications. *Theor. Appl. Genet.*, **124**, 723–38.

Schmid, K.J., Ramos-Onsins, S., Ringys-Beckstein, H., Weisshaar, B. and Mitchell-Olds, T. (2005) A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics*, **169**, 1601–15.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of Arabidopsis thaliana development. *Nat. Genet.*, **37**, 501–6.

Schnurr, J., Shockey, J., Boer, G. De and Browse, J.A. (2002) Fatty acid export from the chloroplast. Molecular characterization of a major plastidial acyl-coenzyme A synthetase from Arabidopsis. *Plant Physiol.*, **129**, 1700–1709.

Sharbel, T.F., Haubold, B. and Mitchell-Olds, T. (2000) Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. *Mol. Ecol.*, **9**, 2109–18.

Sharma, A. and Chauhan, R.S. (2012) In silico identification and comparative genomics of candidate genes involved in biosynthesis and accumulation of seed oil in plants. *Comp. Funct. Genomics*, **2012**, 914843.

Shindo, C., Bernasconi, G. and Hardtke, C.S. (2007) Natural genetic variation in Arabidopsis: tools, traits and prospects for evolutionary ecology. *Ann. Bot.*, **99**, 1043–54.

Shockey, J., Fulda, M. and Browse, J.A. (2002) Arabidopsis contains nine long-chain acyl-coenzyme a synthetase genes that participate in fatty acid and glycerolipid metabolism. *Plant Physiol.*, **129**, 1710–1722.

Ståhl, U., Carlsson, A. and Lenman, M. (2004) Cloning and functional characterization of a phospholipid: diacylglycerol acyltransferase from Arabidopsis. *Plant ...*, **135**, 1324–1335.

Stein, C.M. and Elston, R.C. (2009) Finding genes underlying human disease. *Clin. Genet.*, **75**, 101–6.

Sussman, M., Amasino, R., Young, J.C., Krysan, P.J. and Austin-Phillips, S. (2000) The Arabidopsis knockout facility at the University of Wisconsin–Madison. *Plant*, **124**, 1465–1467.

Thomas (ed.), M. (2012) *Oil World Annual* Thomas Mielke, ed., Hamburg, Germany: ISTA Mielke GmbH.

Vanhercke, T., Wood, C.C., Stymne, S., Singh, S.P. and Green, A.G. (2013) Metabolic engineering of plant oils and waxes for use as industrial feedstocks. *Plant Biotechnol. J.*, **11**, 197–210.

Wallis, J.G. and Browse, J. (2010) Lipid biochemists salute the genome. *Plant J.*, **61**, 1092–106.

Wang, H., Guo, J., Lambert, K.N. and Lin, Y. (2007) Developmental control of Arabidopsis seed oil biosynthesis. *Planta*, **226**, 773–83.

Weigel, D. (2012) Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant Physiol.*, **158**, 2–22.

Wheeler, M., Tronconi, M., Drincovich, M.F., Andreo, C.S., Flugge, U.-I. and Maurino, V.G. (2005) A comprehensive analysis of the NADP-malic enzyme gene family of Arabidopsis. *Plant ...*, **139**, 39–51.

Williams, T.P.H. and P.N. (1964) *The chemical constitution of natural fats.*, Wiley, New York.

Wright, S.I. and Gaut, B.S. (2005) Molecular population genetics and the search for adaptive evolution in plants. *Mol. Biol. Evol.*, **22**, 506–19.

Xu, J., Carlsson, A.S., Francis, T., Zhang, M., Hoffman, T., Giblin, M.E. and Taylor, D.C. (2012) Triacylglycerol synthesis by PDAT1 in the absence of DGAT1 activity is dependent on re-acylation of LPC by LPCAT2. *BMC Plant Biol.*, **12**, 4.

Yang, Q., Fan, C., Guo, Z., Qin, J., Wu, J., Li, Q., Fu, T. and Zhou, Y. (2012) Identification of FAD2 and FAD3 genes in Brassica napus genome and development of allele-specific markers for high oleic and low linolenic acid contents. *Theor. Appl. Genet.*, **125**, 715–29.

Zhang, M., Fan, J., Taylor, D.C. and Ohlrogge, J.B. (2009) DGAT1 and PDAT1 acyltransferases have overlapping functions in Arabidopsis triacylglycerol

biosynthesis and are essential for normal pollen and seed development. *Plant Cell*, **21**, 3885–901.

Zhao, K., Aranzana, M.J., Kim, S., *et al.* (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet.*, **3**, e4.

Zheljazkov, V.D., Vick, B. A., Ebelhar, M.W., Buehring, N., Baldwin, B.S., Astatkie, T. and Miller, J.F. (2008) Yield, Oil Content, and Composition of Sunflower Grown at Multiple Locations in Mississippi. *Agron. J.*, **100**, 635.